DATA SCIENCE TOOLS

Unit-1

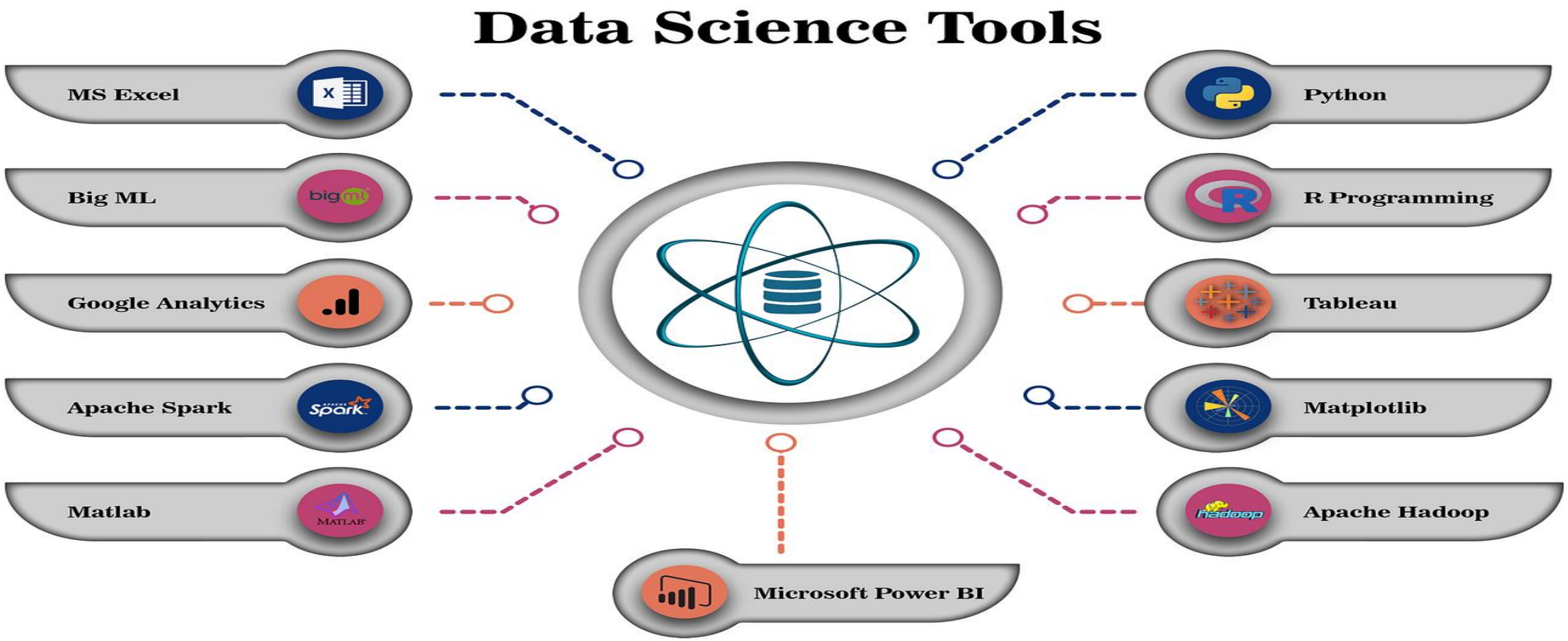
MS Excel

Unit-1: MS Excel (Syllabus)

1. Conditional Formatting
2. Sparkline and Number Formats
3. macros
4. drop down lists
5. Mastering charting techniques
6. Create an Interactive Dashboard

SUB INTRODUCTION:

Data science tools are software applications, frameworks, and libraries designed to facilitate various tasks in the data science workflow, including data collection, preprocessing, analysis, modeling, and visualization. Here's a brief overview of some commonly used data science tools:



❖Programming Languages:

Python: Widely used for its simplicity, versatility, and extensive libraries like Pandas, NumPy, and Scikit-learn.

R: Popular for statistical computing and graphics, with packages like dplyr, ggplot2, and caret.

❖Data Manipulation and Analysis:

Pandas: Python library for data manipulation and analysis, especially suited for structured data.

dplyr: R package for fast data manipulation, featuring a concise syntax.

SQL: Standard language for managing and querying relational databases.

❖Data Visualization:

Matplotlib, Seaborn, Plotly: Python libraries for creating static and interactive visualizations.

ggplot2: R package for producing elegant statistical graphics.

❖Machine Learning:

Scikit-learn: Python library for machine learning algorithms, including classification, regression, clustering, and dimensionality reduction.

TensorFlow, PyTorch: Deep learning frameworks for building neural networks and training models.

❖Big Data Processing:

Apache Spark: Distributed computing framework for processing large-scale data sets.

Hadoop: Framework for distributed storage and processing of big data.

❖Data Wrangling and Transformation:

Apache Airflow: Platform to programmatically author, schedule, and monitor workflows.

Apache NiFi: Data flow automation tool for collecting, transferring, and processing data.

❖Data Integration and ETL:

Apache Kafka: Distributed streaming platform for building real-time data pipelines.

Talend, Informatica: Enterprise-grade tools for data integration and ETL (Extract, Transform, Load) processes.

❖Data Storage:

Relational databases (e.g., MySQL, PostgreSQL) for structured data.

NoSQL databases (e.g., MongoDB, Cassandra) for semi-structured and unstructured data.

Data lakes (e.g., Amazon S3, Azure Data Lake Storage) for storing large volumes of raw data.

❖Notebooks and Development Environments:

Jupyter Notebook, JupyterLab: Interactive environments for data analysis, visualization, and documentation.

RStudio: Integrated development environment (IDE) for R programming.

❖Data Governance and Management:

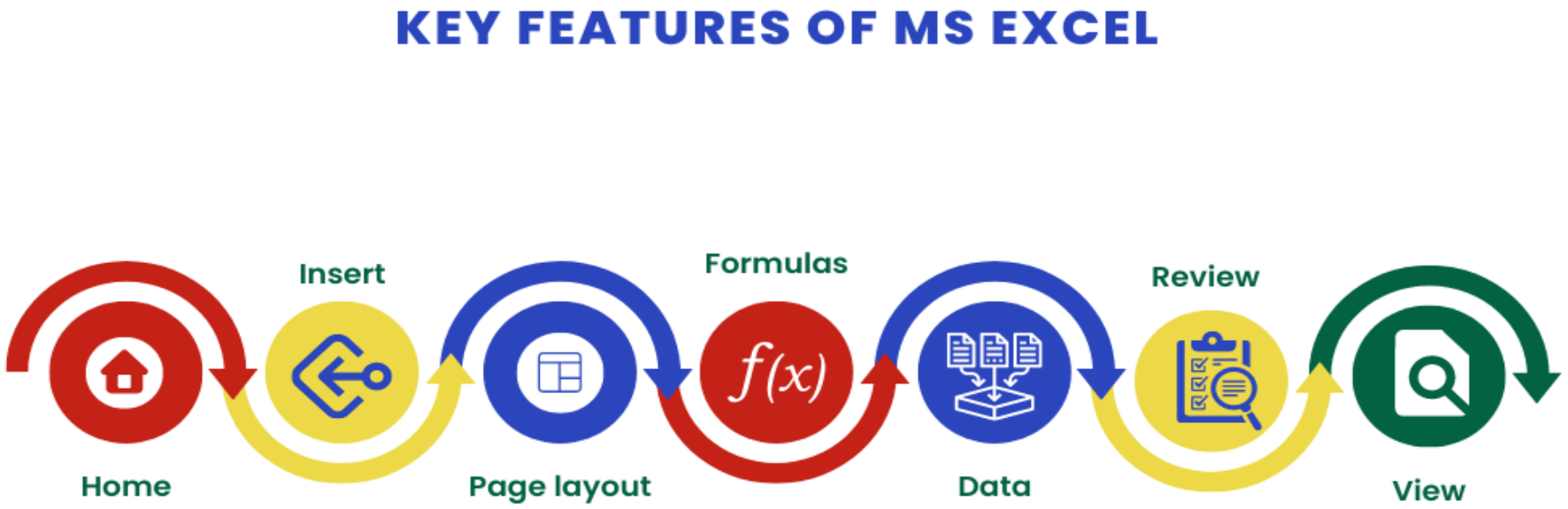
Apache Atlas: Metadata management and governance tool for Hadoop ecosystems.

Collibra: Data governance platform for managing and understanding data assets.

These tools form the foundation of a data scientist's toolkit, enabling them to extract insights from data and build predictive models to solve complex problems across various domains.

Ms Excel

Microsoft Excel is a powerful spreadsheet program used for various purposes such as data organization, analysis, and visualization. Here are some common tasks and features in Excel:



- **Data Entry:** Excel provides a grid-like interface where you can enter data into cells. You can type text, numbers, dates, and formulas into cells.
- **Formulas and Functions:** Excel has a wide range of built-in functions for performing calculations. These functions include SUM, AVERAGE, IF, VLOOKUP, and many more. Formulas can be entered into cells to perform calculations using these functions.

- **Formatting:** You can format cells to change their appearance, such as font style, size, color, alignment, and borders. Conditional formatting allows you to automatically apply formatting based on specific conditions.
- **Charts and Graphs:** Excel enables you to create various types of charts and graphs to visualize your data, including bar charts, line charts, pie charts, scatter plots, and more.
- **Data Analysis:** Excel provides tools for data analysis, such as sorting, filtering, pivot tables, and data validation. These features help you analyze and manipulate large datasets efficiently.
- **Data Import and Export:** You can import data from external sources such as text files, databases, and other Excel files. Excel also supports exporting data to different file formats.
- **Collaboration:** Excel offers collaboration features, such as sharing workbooks with others, tracking changes, and commenting on cells.
- **Macros:** Excel allows you to automate repetitive tasks using macros. You can record a series of actions and play them back as needed.
- **Data Protection:** Excel provides features for protecting your data, such as password protection, cell locking, and worksheet protection.
- **Customization:** You can customize Excel by adding or removing features, creating custom functions, and designing templates.

1. Conditional Formatting

Conditional Formatting is a feature in an Excel spreadsheet. It is used to maintain the status of the result easily. It is most often used as color-based formatting to highlight, emphasize, or differentiate among data and information stored in an Excel spreadsheet.

When it comes to applying alternative forms to data that fit particular criteria, Excel conditional formatting is a highly useful feature. It can make it easier for you to draw attention to the key details in your spreadsheets and quickly identify differences in cell values.

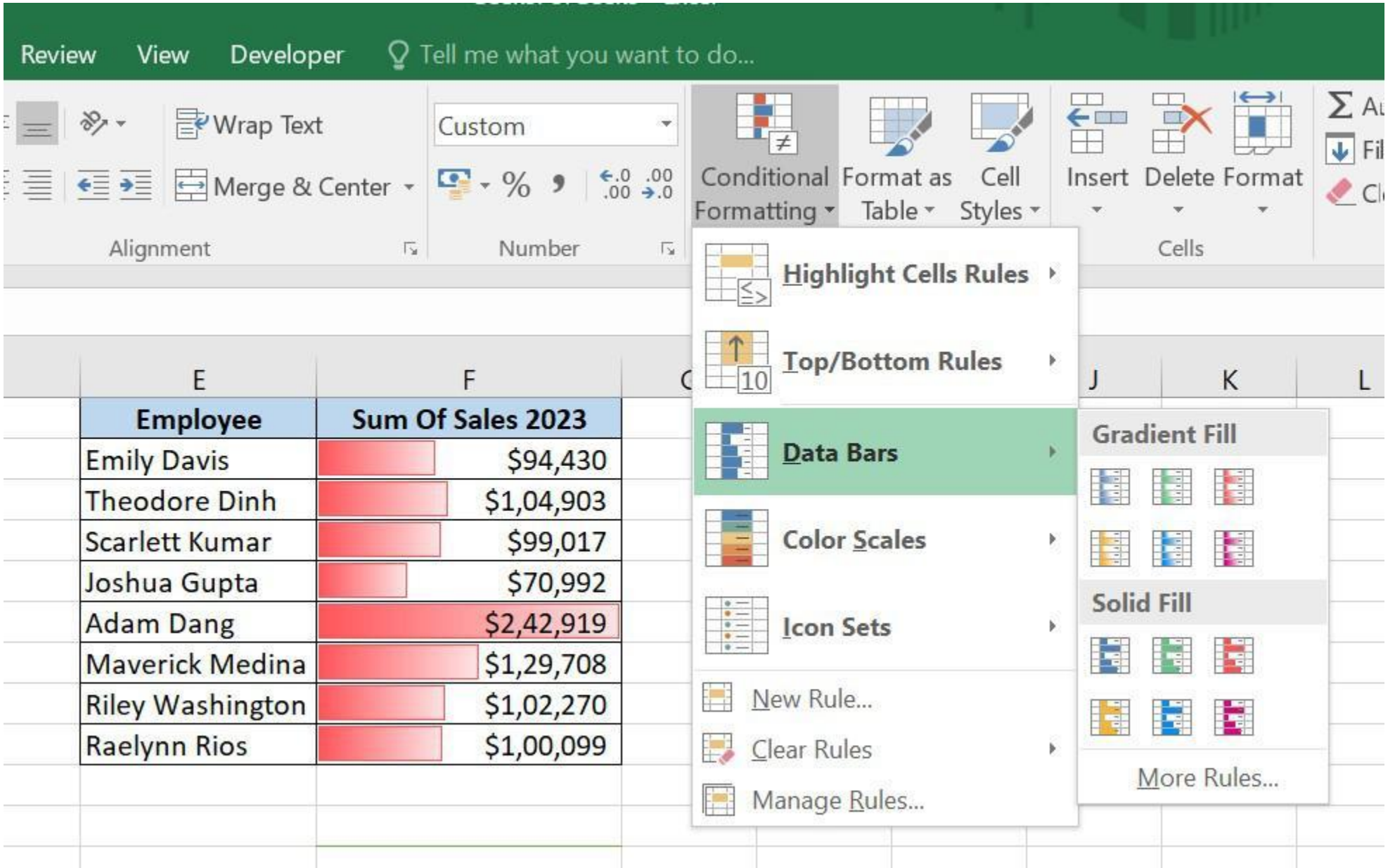
Types of Conditional Formatting Visualization

You can use preset rules, including **Color scales, Data Bars, Icon Sets, Sort filters**, etc. to conditionally format your data, or you can construct your own rules that specify when and how the selected cells should be highlighted.

1. **Color scales**
2. **Data Bars**
3. **Icon Sets**
4. **Sort filters**

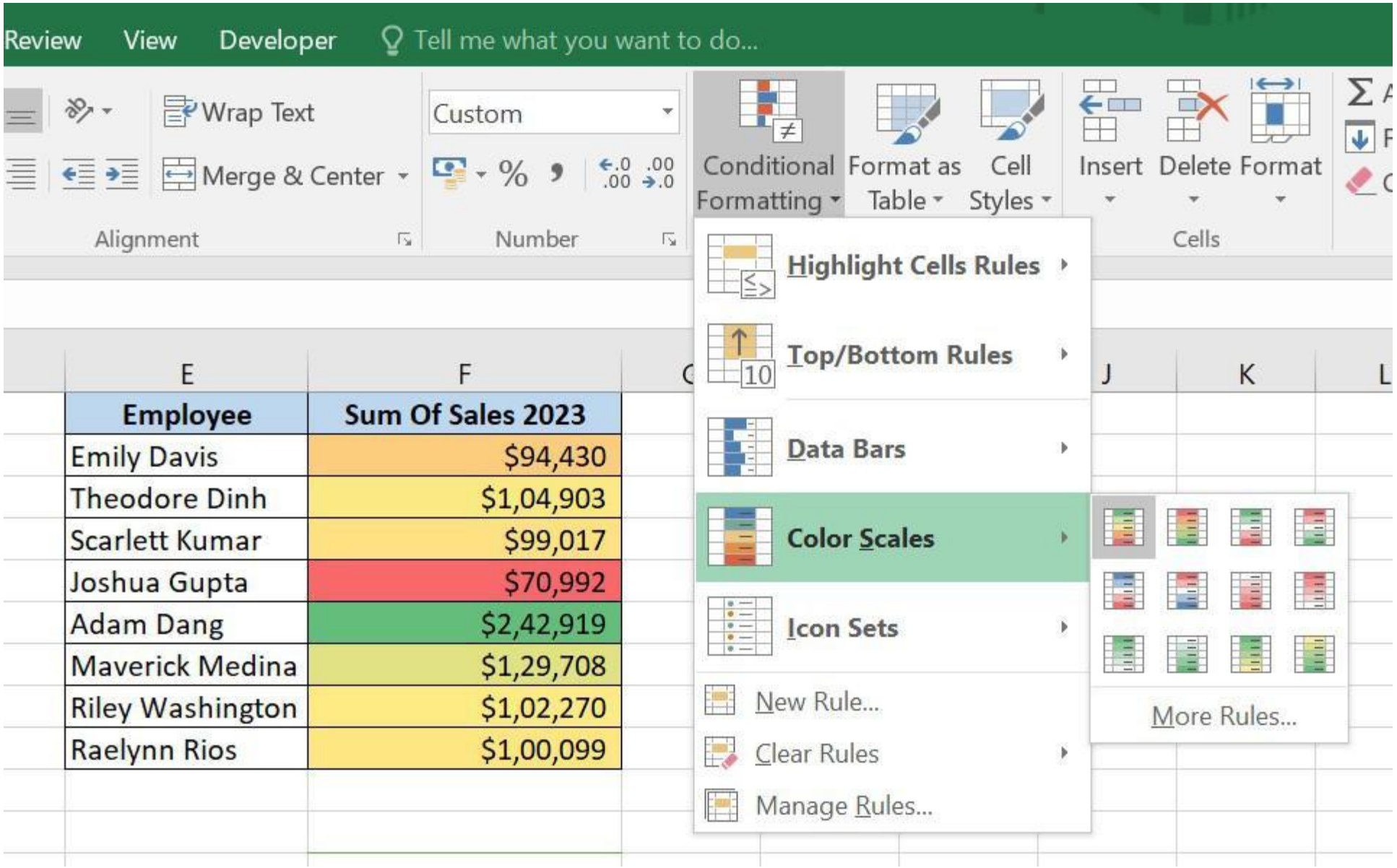
Data Bars

Similar to a bar graph, data bars are horizontal bars that are added to each cell



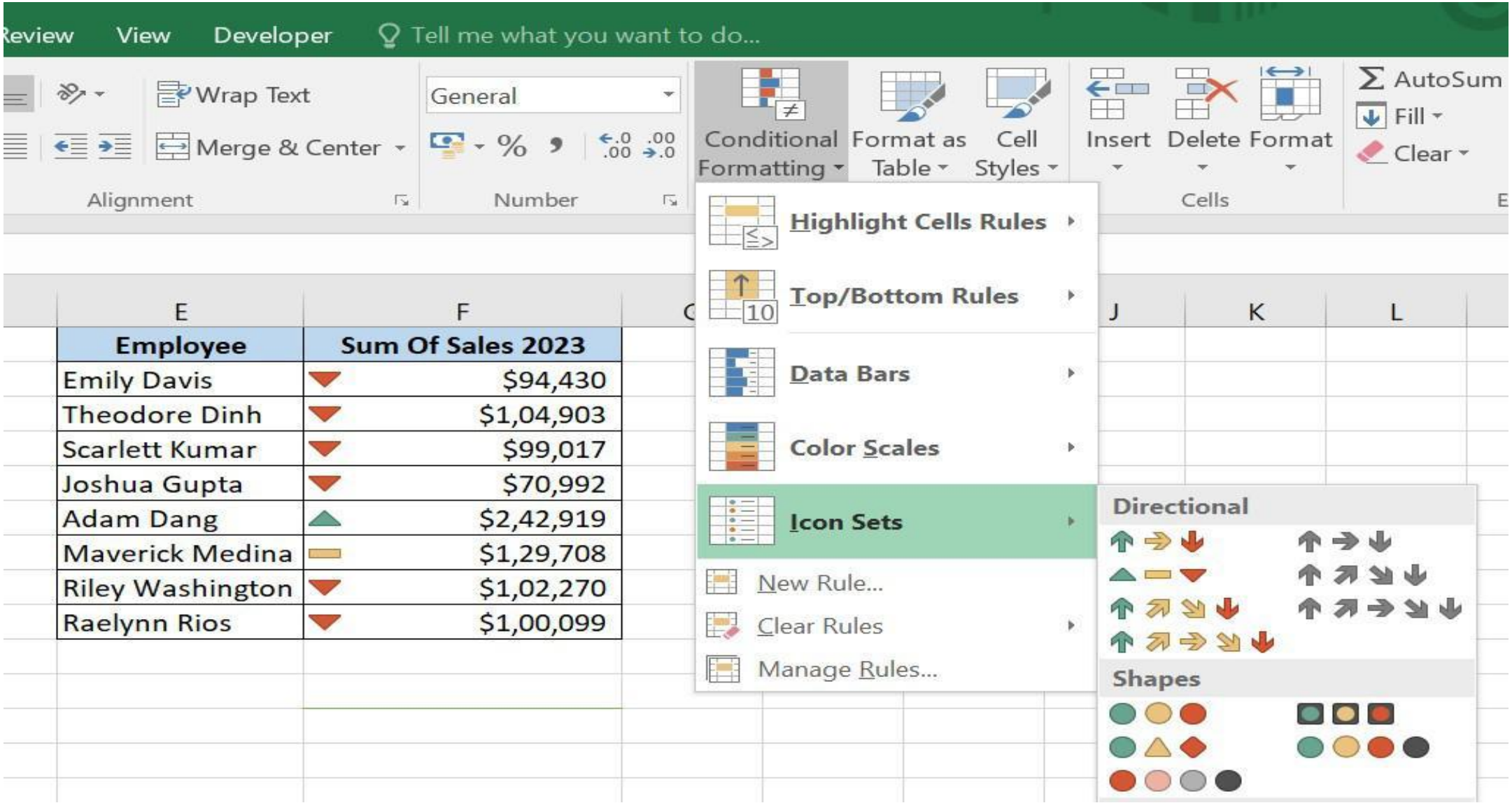
Color scales

Alter each cell's color depending on its value. A two- or three-color gradient is used for each color scale



Icon Sets

Based on the value of each cell, icon sets assign a distinct icon to each cell.



How to Use Conditional Formatting in Excel

Step 1: Select the Cells

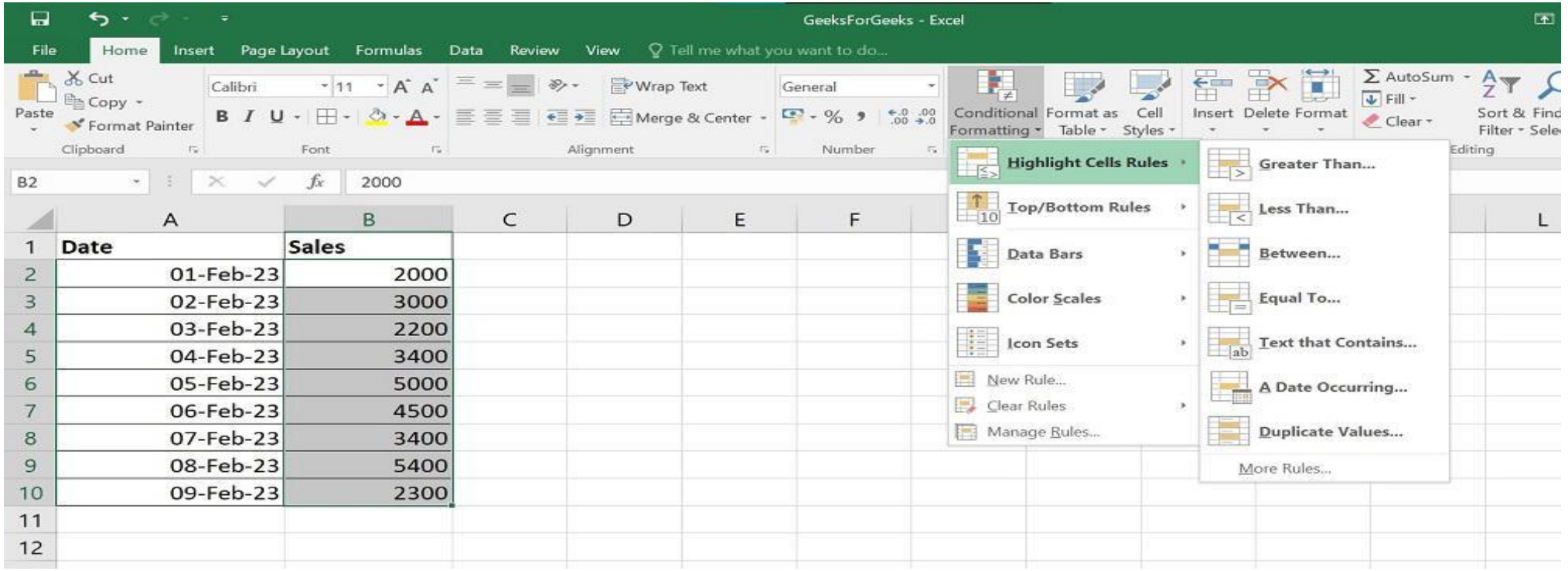
Select the cells you want to format in the Spreadsheet

Step 2: Click on Conditional Formatting

Click on Conditional Formatting, On the Home tab, in the Styles group

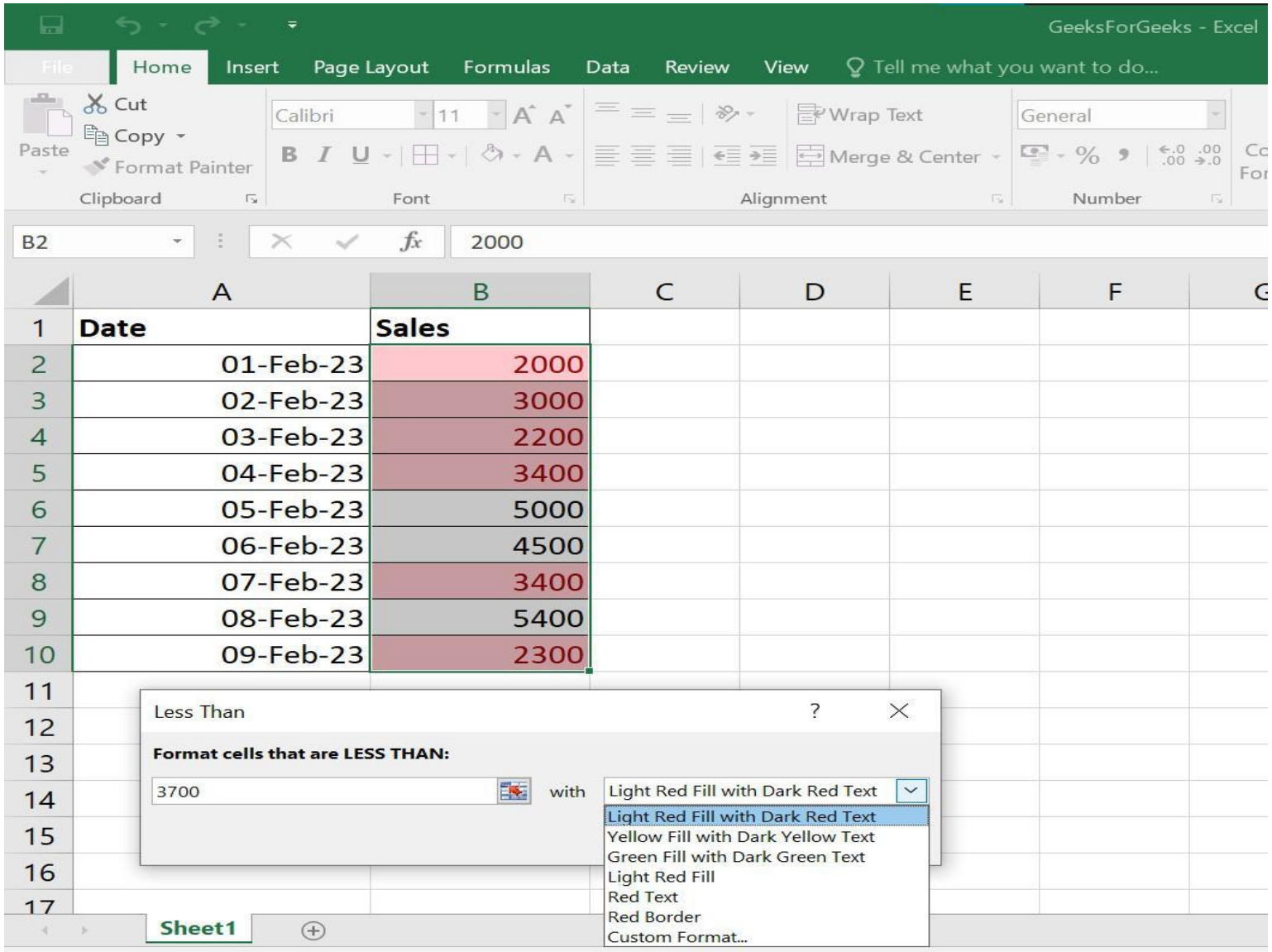
Step 3: From a Set of Preset Rules, Pick the Required One

From a set of preset rules, pick the one that best serves your needs



Step 4: Enter the Value and Select the Chosen Format from the Drop-Down List

Enter the value in the box on the left of the dialogue box and select the chosen format from the drop-down list on the right (with Default Light Red Fill with Dark Red Text)



The flexibility to choose a different rule type that better suits your data. For instance, you can opt for:

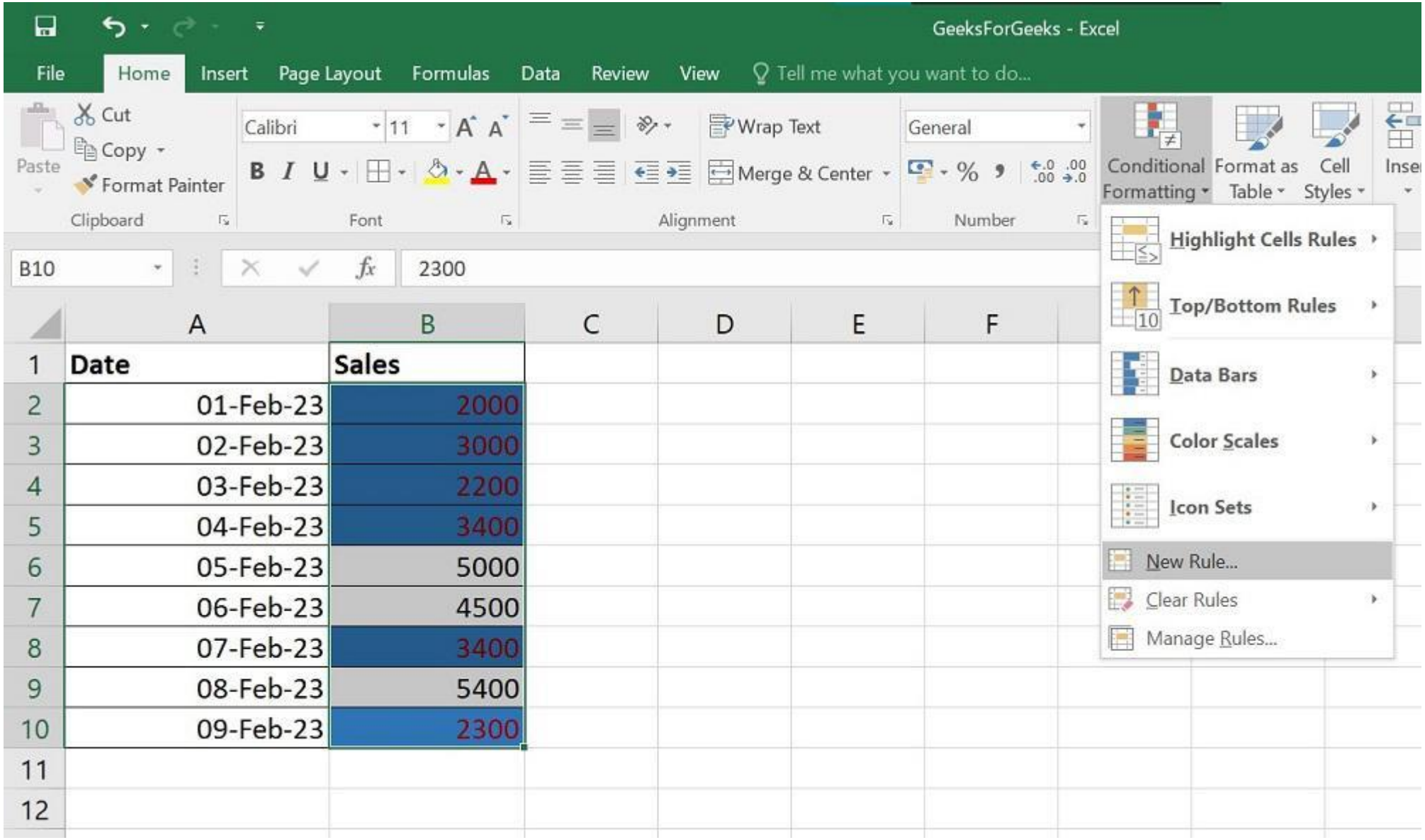
- Greater than or equal to
- Between two values
- Text that contains specific words or characters
- Date occurring in a certain range
- Duplicate values

How to Create a New Conditional Formatting Rule in Excel

Step 1: Select the Cell or Cells you Want to Format

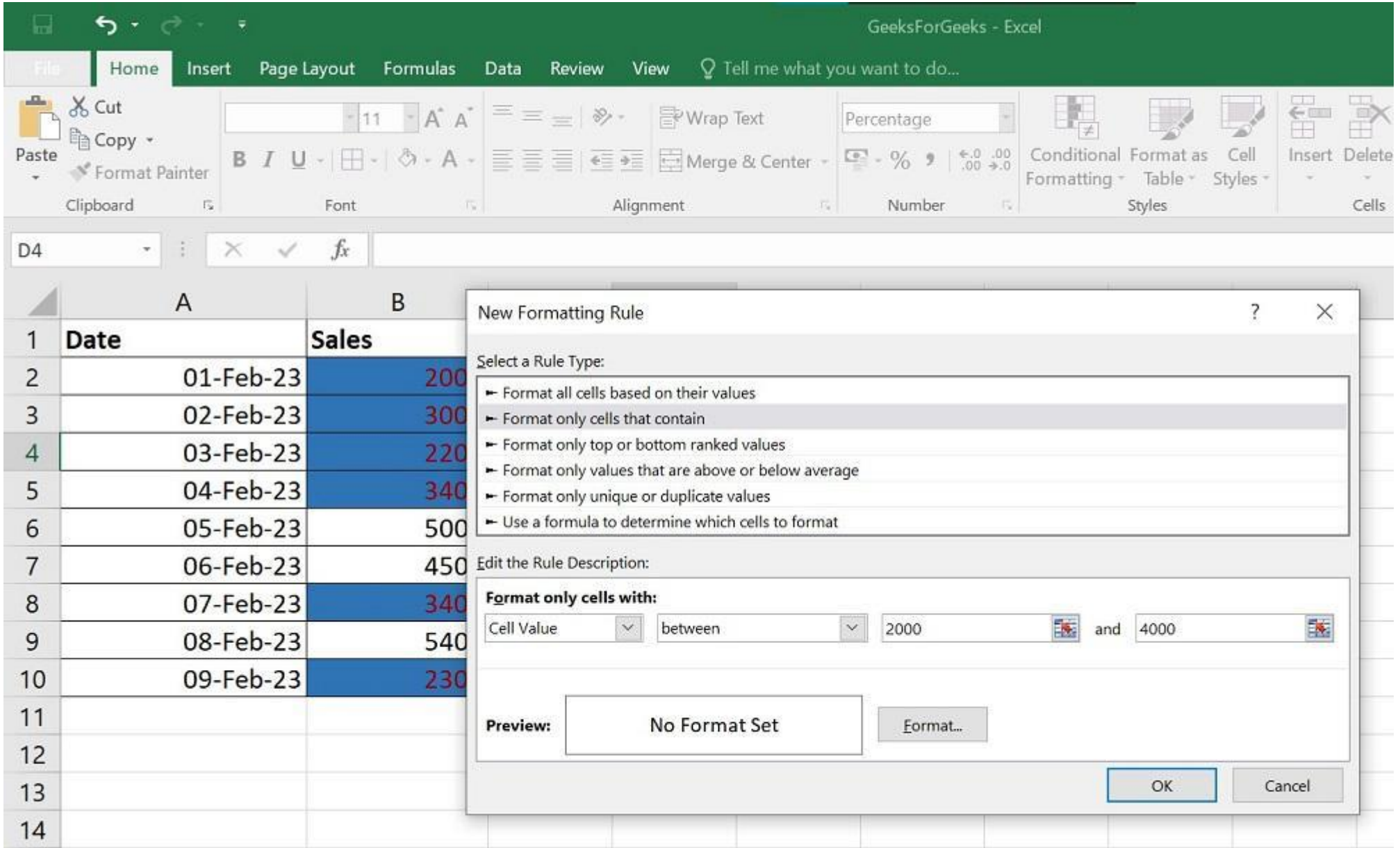
Step 2: Navigate to the Home Tab and Select Conditional Formatting and Click New Rule

New Conditional Formatting Rule in Excel

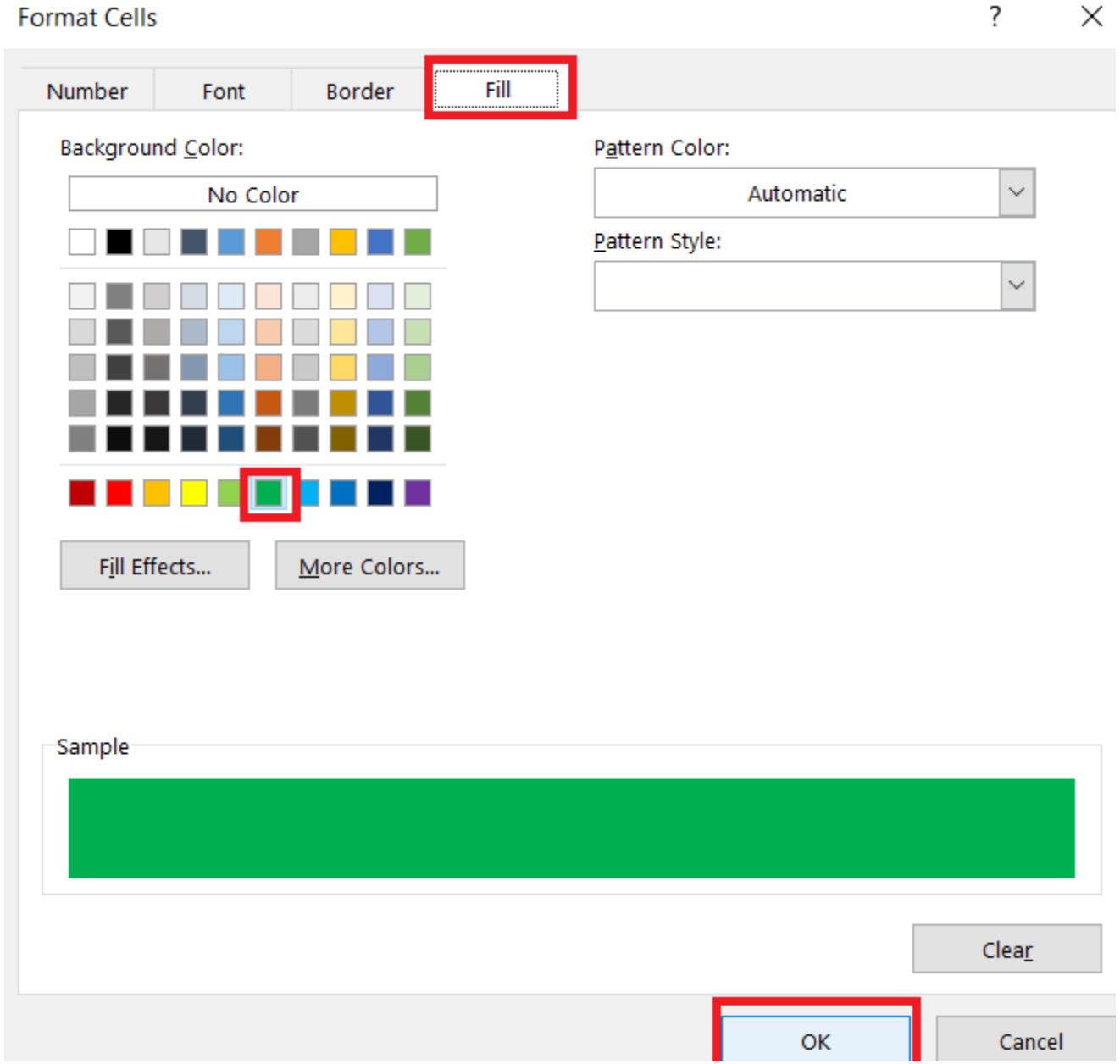


Step 3: Select the Rule Type and Click Ok

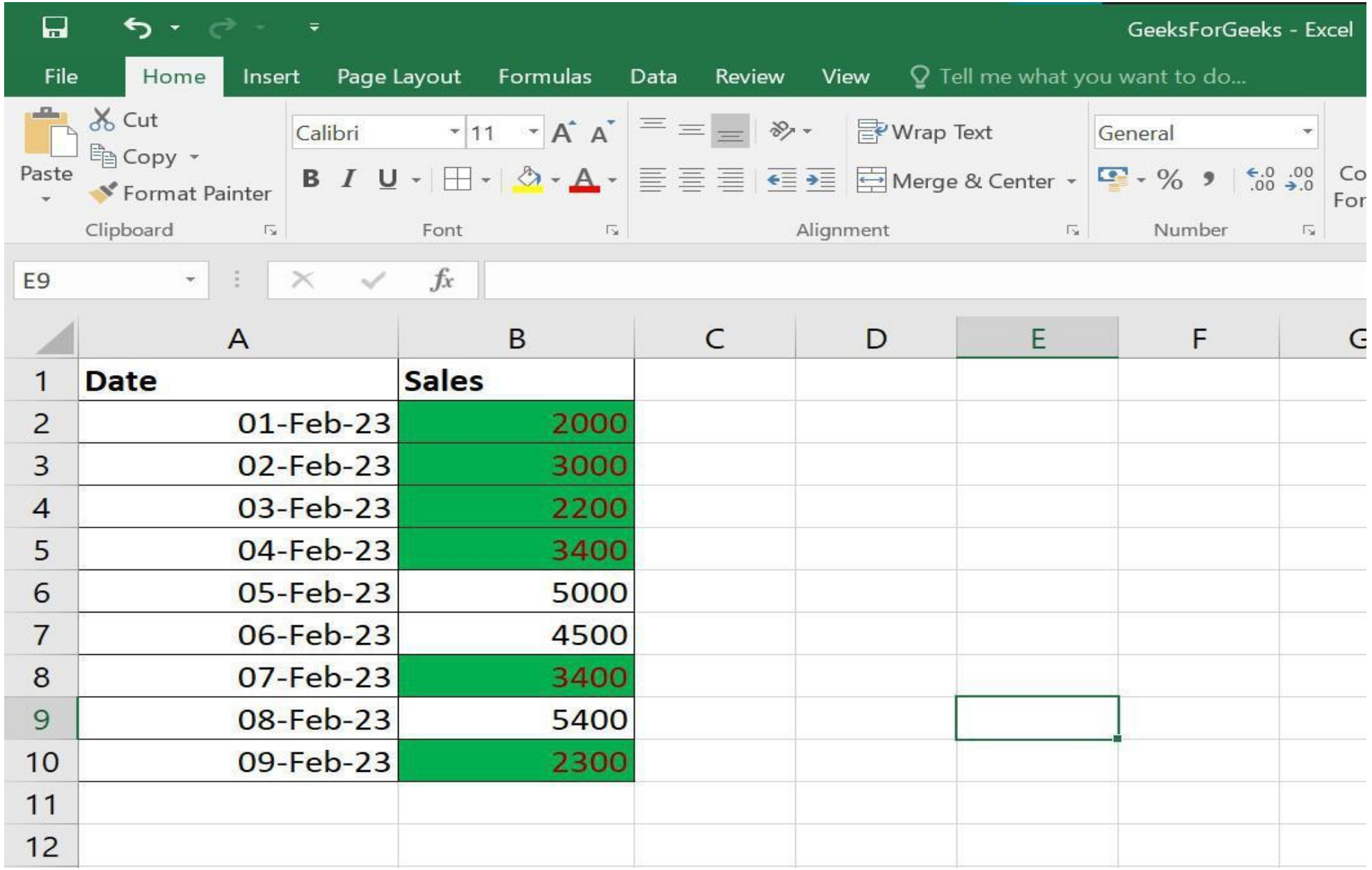
Now select the rule type, in the New Formatting Rule dialog box



Step 4: Click the Format button, Choose the Color, and Click Ok



Step 5: Preview Result



	A	B	C	D	E	F	G
1	**Date**	**Sales**					
2	01-Feb-23	2000					
3	02-Feb-23	3000					
4	03-Feb-23	2200					
5	04-Feb-23	3400					
6	05-Feb-23	5000					
7	06-Feb-23	4500					
8	07-Feb-23	3400					
9	08-Feb-23	5400					
10	09-Feb-23	2300					
11							
12							

How to Edit Conditional Formatting Rules in Excel

To make any changes in an already existing Formatting Rule in Excel, followthese steps

Step 1: Select the Cells

Select any cell to which a rule already applies.

Step 2: Navigate to the Home Tab and Click on Manage Rules under Conditional Formatting

Step 3: Select the Rule you Want to Edit

This will open the Rules Manager dialog box, select the rule you want to modify, then Click on Edit Rule.

Step 4: Make the Changes and Click Ok

Make the required changes in the Edit Formatting Rule dialogue window, Click on Ok

How to Highlight Cells with Value Greater or Less than a Number in Excel

Follow the below steps for highlighting cells with a value greater or less than a given number:

Step 1: Select the Cells

Select the cells that you want to highlight with Greater or less than a number.

Step 2: Navigate to the Home Tab and Click on Conditional Formatting

Click on the conditional formatting and click on new rules. Now select “Format all cells based on their values“. In the ‘Minimum’ and choose the value you want to identify the lesser values and choose the color. Similarly, In the Maximum and choose the value you want to identify the lesser values and choose the color. And click ok.

Step 3: Preview Result

Excel Conditional Formatting Best Uses

- **Highlighting Top and Bottom Values:** You can use Conditional Formatting to highlight the highest and lowest values in a Dataset so that you can easily identify significant data points.
- **Data Bars and Colors Scales:** Represent data visually using data bars or color scales to create intuitive maps, making patterns and trends more apparent.
- **Icons Sets:** Apply Icon Sets to cells, by using symbols such as arrows or checkmarks, to quickly assess data against preset conditions.

Advanced Conditional Formatting Techniques

Formulas in Conditional Formatting: You can use custom formulas to create complex conditions for formatting. This can help you for more precise control over the formatting rules.

Cell Reference in Conditional Formatting: You can apply Conditional Formatting dynamically based on values in other cells using Cell References.

Managing Rules: You can efficiently Manage and organize your conditional formatting rules through the “Conditional Formatting Rules Manager”.

2) Sparkline and Number Formats:

What are Sparklines

Sparklines are miniature, condensed charts that provide a quick visual representation of data trends and patterns within a single cell in Microsoft Excel. These tiny graphs allow you to analyze data at a glance, without the need to create full-sized charts. Sparklines are an excellent tool for displaying data in a compact and easy-to-read format, making them particularly useful for dashboards, reports, and presentations.

In this comprehensive guide, we will explore the concepts of sparklines, how to create them in Excel, and various customization options to enhance their effectiveness. Unlike regular charts, sparklines are not objects. These reside in a cell as the background of that cell.

Types of Sparklines in Excel

There are three types of sparklines in Excel:

- ❖Line
- ❖Column
- ❖Win-Loss

How to Insert Sparklines in Excel

Follow the below steps to add sparkline in Excel:

Step 1: Go to the Insert tab.

Step 2: Select the cell where you want to insert the sparkline

>Line” width=”inherit” height=”inherit”>

Go to insert > Sparklines > Select option

Step 3: Select the range of cells for which you have to add a sparkline.

Step 4: Click on the OK button, the sparkline is now added to the selected cell.

Customize the Sparkline

To format the sparkline, select the sparkline you want to format and then select the Sparkline option on the menubar

Using show submenu

In the show submenu, there are 6 options :

1. **High Point:** Marks the maximum point on the sparkline.
2. **Low Point:** Marks the minimum point on the sparkline.
3. **First Point:** Marks the leftmost (first) point on the sparkline.
4. **Last Point:** Marks the rightmost (last) point on the sparkline.
5. **Markers:** Mark all the edge points on the sparkline.
6. **Negative Points**: Marks the negative points in the graph with a different color

Using Style submenu

In the style submenu, we can select the style of the sparkline, many options are available (different variety of colors). We can also set the color of the sparkline and marker color using Sparkline Color and Marker Color options.

Using Group submenu

In the Axis section, there are many options by which you can reformat your sparkline. One of the options is Plot Data Right-to-Left, in this case, the data is plotted on the sparkline from the rightmost data to the leftmost data.

Group, Ungroup, and Clear Sparklines

- **Group:** It groups all the sparklines together. In simple words, you can treat all the sparklines as having the same characteristics and applying style on any one of them, styles all the other sparklines in that group. If we change the color of one sparkline, all the other sparklines get affected.
- **Ungroup**: Ungroup option reverses the effects done by the Group option, it again makes the sparklines as different entities.
- **Clear**: This option can clear/delete the selected sparkline or the selected sparkline group(all the sparklines in that group).

Edit the DataSet of Existing Sparklines

Editing the dataset of existing sparklines in Excel is a simple process that allows you to modify the data and update the sparklines accordingly, Follow the below steps to edit the dataset:

Step 1: Select the Sparklines

First, click on the sparkling that you want to edit. Excel will highlight the selected sparkling, including that it is active for editing.

Step 2: Access the Edit Data Option

Next, go to the “Design” tab in the Excel ribbon. In the “Type” group, you will find the “Edit Data” drop-down menu.

Step 3: Choose the Editing Option

Click on the “Edit Data” drop-down menu to reveal the following options:

Edit Group Location & Data

Select this option when you have grouped multiple sparklines, and you want to change the data for the entire group. Grouping Sparklines is a useful feature for managing and updating multiple Sparklines simultaneously.

Edit Single Sparkline’s Data

Choose this option to change the data for the selected sparkling only. This is helpful when you have individual sparklines with unique data that you want to modify independently

Hidden and Empty Cells

When working with the line sparklines in a dataset that contains empty cells, you may observe that the sparkline displays a gap to indicate the absence of data in those cells. This gap can sometimes disrupt the continuity o the sparkline and affect data visualization. However, Excel provides effective ways to handle hidden and empty cells within sparklines, ensuring a smooth and acute representation of data trends.

Ignoring Hidden Cells

By default, Excel considers hidden cells in sparklines, leading to gaps in the line when data is hidden. To overcome this, you can instruct Excel to ignore hidden cells while creating sparklines. Follow the below steps to ignore the hidden cells:

Step 1: Click the cell that has the sparkline.

Step 2: Click the Design Tab.

Step 3: Click on the Edit data Option.

Step 4: In the drop-down, select the ‘Hidden & Empty Cells’ option.

How to change sparklines in Excel

You can also change the Sparkline type – form line to column or vice-versa, you can do this using the following steps:

Step 1: Click the sparkling you want to change.

Step 2: Click the sparkling Tools Design Tab.

Step 3: In the type group, select the sparkling you want.

Highlighting Data Points in Sparklines

some different highlighting options available :

- High/Low Point: This option is used to highlight the maximum and/or minimum data points.
- First/Last Point: This is used to highlight the First/ last data points.
- Negative Points: This option can be used to highlight all negative points if present in your data.
- Markers: Exclusive to line sparklines, this option highlights all data points with a marker. You can customize the marker color using the ‘Marker Color” option.

Number Format :

MS Excel or Microsoft Excel is powerful spreadsheet software developed by Microsoft by combining all the necessary features, functions, and tools. When it comes to formatting data within a spreadsheet, Excel has several formatting options to format different types of values accordingly.

One of the essential features in Excel is to use number format, which enables us to format any numerical value entered within cells. Excel has a wide range of built-in number formats. As a bonus, it also supports custom number formats.

What is Number Format?

A number format in Excel refers to a specific code or format that controls the way values are displayed within Excel cells. Excel has several inbuilt formats to help us display values in different ways in spreadsheets for Number, Percentage, Currency, Accounting, Date and Time, etc.

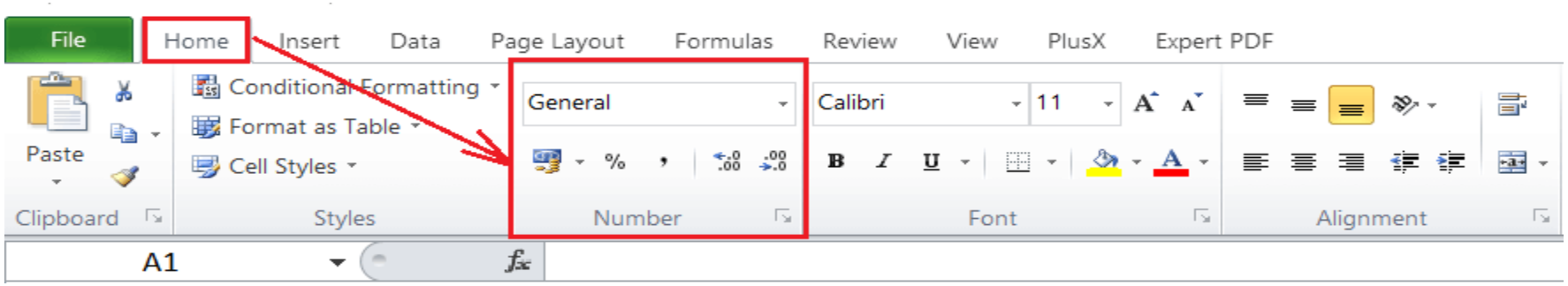
Example: The following table displays different formats of the same date based on the given or applied code:

Given Date in Cell	Applied Code	Result/ Value displayed in Cell
1-Jan-2021	yyyy	2021
1-Jan-2021	yy	21
1-Jan-2021	mmm	Jan
1-Jan-2021	mmmm	January
1-Jan-2021	d	1
1-Jan-2021	ddd	Fri
1-Jan-2021	dddd	Friday

The main purpose of the number format is to change the way numerical values are displayed. This does not affect the actual value in the cell, and we can still do the relevant calculations accordingly. In short, the underlying value recorded within the cell is not changed.

How to access Number Formats in Excel?

To access the number formats in Excel, we need to navigate the Home tab and locate the 'Number' group. It looks like this:



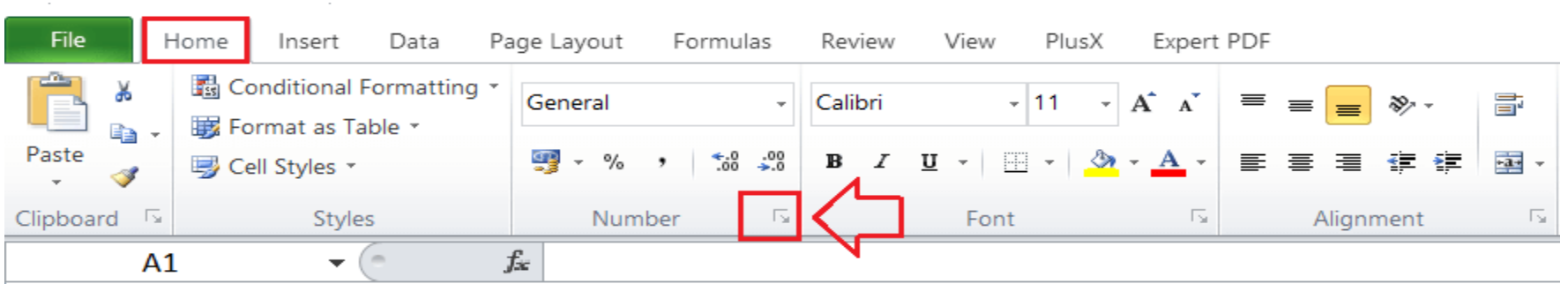
In the above image, these are the built-in number formats. We can click the drop-down icon next to 'General' to access all the common number formats. However, we need to go to the Format Cells dialogue box to access more number formats.

How to create a custom number format in Excel?

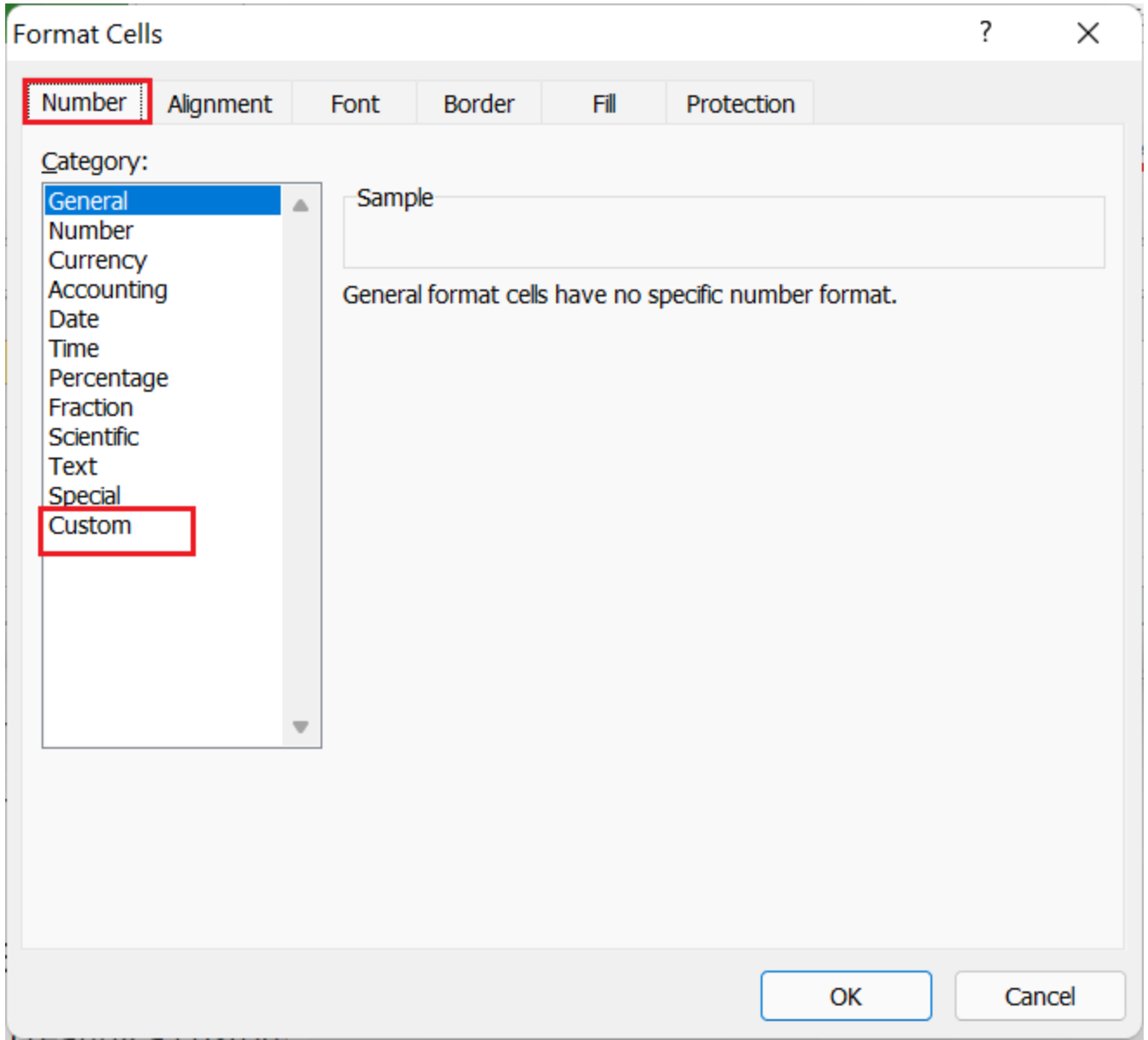
To access all the number formats in Excel or to create custom number formats, we need to follow the below steps:

First, we need to select or highlight all the cells we are willing to apply the number format or custom number format.

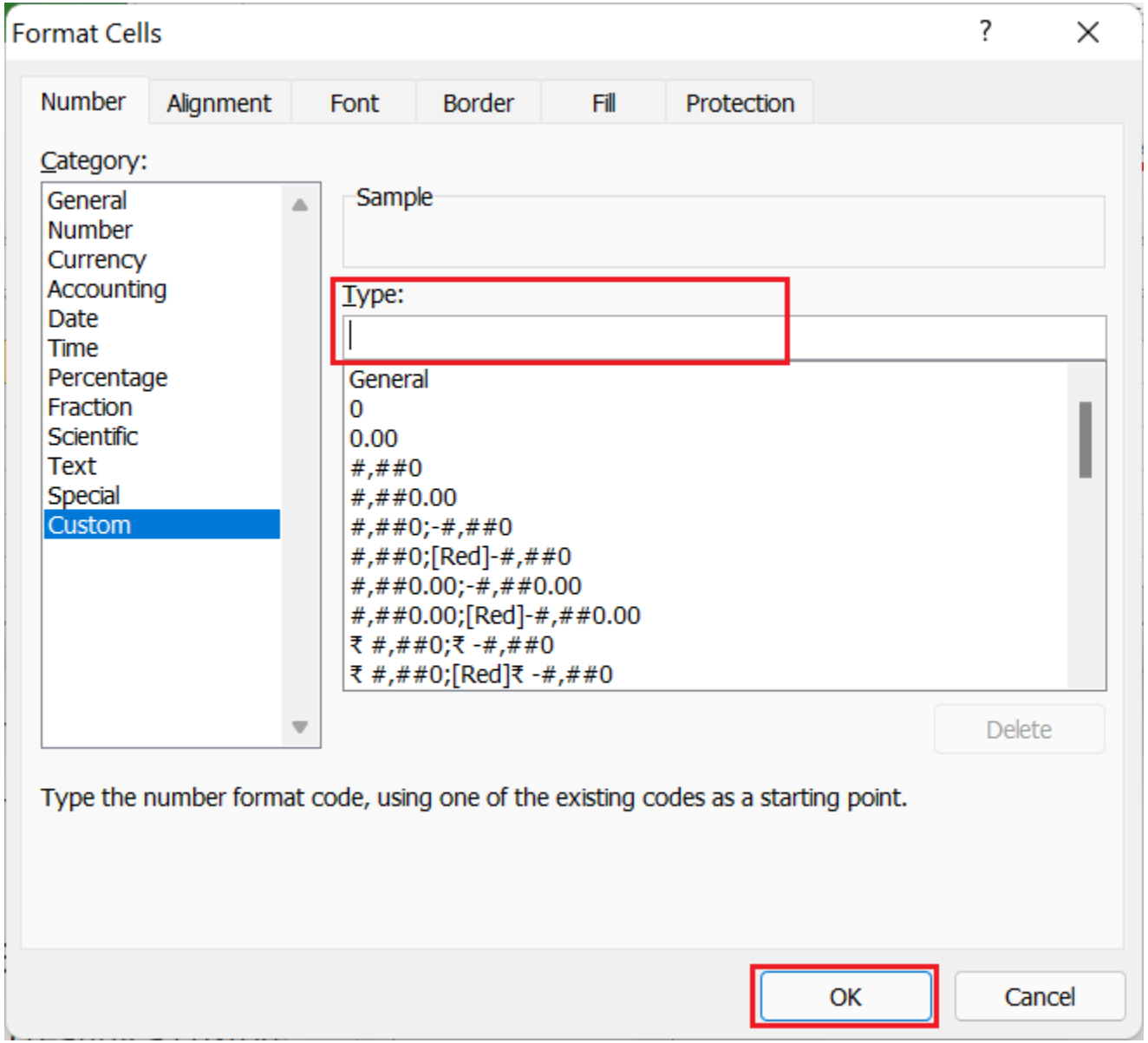
Next, we need to launch the Format Cells dialogue box. For this, we need to click on the More button from the Home > Number group, as shown below:



In the Format Cells dialogue box, we need to select the Number tab if not selected. Under the category option, we can access all the available number formats of Excel. We must click the Custom option to create our desired number format.



In the next window, we need to enter the format code under the Type box, and it will be applied to selected cells. The format code must be typed with proper structure.



Important Rules for Custom Number Format

- To create a custom number format without errors, we must remember the following rules:
- If we enter only one section in custom number format, the format is applied to all the number types, such as positive, negative, and zeros.
- If we enter two sections in custom number format, the first section is applied to positive numbers and zeros, while the second section is applied to negative numbers.
- A custom number format must contain all four sections to apply the given format to text values.
- If we want to use the default number format for any middle section, we can type **General** instead of entering the corresponding format in our custom number format. For example: *General; -General; "-"; General.* Here, we have added the minus sign in the second section as the second section denotes the formatting for the negative numbers.
- Excel also allows us to edit the built-in Excel format. However, the copy of the edited format is created while the original number format remains intact from the customizations.
- To hide any desired section or value type(s), we can skip the specific section in custom format by typing only the ending semicolon. For example- *General; ; ; General*. The negative value and zeros will only be displayed in the formula bar, not within the cell area.

In Microsoft Excel, there are several number formats available to customize the appearance of numerical data. Here's a brief overview:

1. **General**: This is the default format for numbers in Excel. It displays numbers exactly as they are entered.
2. **Number**: This format is used for general display of numbers and can include commas as thousand separators and control the number of decimal places.
3. **Currency**: Formats numbers as currency, typically with a currency symbol, comma separators, and a specified number of decimal places.
4. **Accounting**: Similar to the currency format but aligns currency symbols and decimal points in a column, making it easier to read financial data.
5. **Date**: Formats numbers as dates, allowing for various date formats depending on regional settings.
6. **Time**: Formats numbers as time, displaying hours, minutes, and seconds.
7. **Percentage**: Displays numbers as percentages, multiplying the cell value by 100 and adding a percentage sign.

8.**Fraction**: Converts numbers to fractional form, allowing customization of the denominator.

9.**Scientific**: Displays numbers in scientific notation, useful for very large or very small numbers.

10.**Text**: Treats the cell contents as text, regardless of whether it's a number, date, or other data type.

11.**Custom**: Allows for customization of number formats using specific codes to control how numbers are displayed, including adding text, symbols, and conditional formatting.

3. Macros

Macros in Excel are written in VBA (Visual Basic for Applications) and can be used to automate repetitive tasks, perform complex calculations, interact with users, and much more. Here's a basic guide to creating and running macros in Excel:

1. **Enable the Developer Tab:**

If you haven't already, you'll need to enable the Developer tab in Excel. Go to File > Options > Customize Ribbon, then check the box next to "Developer" in the right-hand column.

2. **Open the Visual Basic Editor:**

Click on the "Developer" tab, then click on "Visual Basic" in the "Code" group. Alternatively, you can press Alt + F11.

3. **Create a New Macro:**

Once the Visual Basic Editor is open, you can insert a new module by right-clicking on the "VBAProject" for your workbook in the Project Explorer window, selecting Insert > Module. Double-click the new module to open it and start writing your macro.

4. **Write Your Macro:**

Write your VBA code in the module. Here's an example of a simple macro that formats the selected cells as bold:

```
Sub MakeBold()
    Selection.Font.Bold = True
End Sub
```

5.Run Your Macro:

You can run your macro from the Developer tab by clicking on "Macros" in the "Code" group, selecting the macro you want to run, and clicking "Run". Alternatively, you can assign the macro to a button or a keyboard shortcut.

6.Assign a Macro to a Button:

You can add a button to your Excel worksheet and assign your macro to it. To do this, go to the Developer tab, click on "Insert" in the Controls group, and then click on "Button". Draw the button on your worksheet, and then assign your macro to it by selecting it from the list.

7. Assign a Macro to a Keyboard Shortcut:

You can also assign a keyboard shortcut to your macro. To do this, go to the Developer tab, click on "Macros" in the Code group, select your macro, and click on "Options...". Here, you can type a letter or number to assign as a shortcut key.

4. drop down lists

1.Create the List:

First, create the list of items you want to appear in the drop-down menu. You can do this in a separate column or on another sheet.

2.Select the Cell:

Select the cell or cells where you want the drop-down list to appear.

3.Open Data Validation:

Go to the "Data" tab on the Excel ribbon, and click on "Data Validation" in the "Data Tools" group.

4. Choose the Validation Criteria:

In the Data Validation dialog box, select "List" from the drop-down menu in the "Allow" field.

5. Specify the Source:

In the "Source" field, either type in the range of cells that contain your list of items or click the button at the end of the field and select the cells containing your list.

6. Optional Settings: You can also set other options such as error alert messages or input messages if desired.

7.Click OK: Once you've set up the validation criteria, click "OK" to apply the drop-down list to the selected cell(s).

Now, when you click on the cell with the drop-down list, you should see a small arrow appear. Clicking on this arrow will display the items from your list, and you can select one of them.

5. Mastering charting techniques

Mastering charting techniques in Microsoft Excel involves understanding how to effectively represent data visually using various chart types and customizing them to convey insights clearly. Here's a clear step-by-step guide:

1. **Selecting Data:** First, choose the data you want to visualize. Make sure it's organized properly with headers and consistent formatting.

2. **Inserting a Chart:** Select the data range, then go to the "Insert" tab on the Excel ribbon. Choose the type of chart that best suits your data from the "Charts" group. Common types include column, bar, line, pie, and scatter plots.

3. **Choosing the Right Chart Type:** Different types of data call for different chart types. For example:

- ❖Use column or bar charts for comparing values across categories.
- ❖Line charts are suitable for showing trends over time.
- ❖Pie charts display parts of a whole.
- ❖Scatter plots visualize the relationship between two sets of data.

4.**Customizing the Chart:**

❖Titles and Labels: Add a title that clearly describes the chart. Label axes appropriately.

❖Formatting: Change colors, fonts, and styles to make the chart visually appealing and easy to read.

❖Gridlines and Axes: Adjust gridlines and axes scales to highlight important data points.

❖Legend: If applicable, include a legend to explain data series or categories.

5. **Adding Data Labels and Data Tables:** Data labels can be added to individual data points or series to provide additional context. Data tables display the underlying data for reference.

6. **Using Chart Tools:** Excel offers various tools to enhance and refine charts:

❖Chart Styles: Apply predefined styles to quickly change the appearance of the chart.

❖Chart Elements: Customize specific elements like data labels, axes, and legends.

❖Chart Filters: Filter data directly within the chart to focus on specific subsets of data.

7. **Utilizing Advanced Features:**

❖Trendlines: Add trendlines to show patterns or forecast future trends.

❖Combination Charts: Combine different chart types in a single chart for better comparison.

❖Secondary Axes: Use secondary axes to compare datasets with different scales.

8.**Updating and Refreshing Charts:** If your data changes, update the chart to reflect the latest information. Excel also allows you to set up dynamic charts that automatically update when new data is added.

9. **Practice and Experiment:** The best way to master charting techniques is through practice. Experiment with different chart types, styles, and features to find what works best for your data and audience.

6.Create an Interactive Dashboard

Creating an interactive dashboard in Microsoft Excel can be achieved using various features like PivotTables, slicers, charts, and form controls. Here's a step-by-step guide to creating a simple interactive dashboard:

1. Data Preparation:

- Prepare your data in Excel. Ensure it's organized and clean.
- Add column headers and format your data as a table (optional but recommended for dynamic range).

2 .Create PivotTables:

- Go to the "Insert" tab and click on "PivotTable."
- Select the range of data you want to analyze and choose where you want the PivotTable to be placed (either a new worksheet or existing worksheet).
- Drag and drop fields from the Field List to the Rows, Columns, and Values areas to structure your PivotTable.

3.Create Charts:

- Highlight the data you want to include in your chart.
- Go to the "Insert" tab and select the type of chart you want (e.g., bar chart, line chart, pie chart).
- Customize your chart as needed (titles, axis labels, formatting).

4. Add Slicers:

- Select your PivotTable.
- Go to the "Options" tab, and click on "Insert Slicer."
- Choose the fields you want to use as slicers (fields you want to filter data by).
- Arrange the slicers on your dashboard.

5. Create Interactive Elements:

- Go to the "Developer" tab (if you don't see it, enable it in Excel options).
- Click on "Insert" in the Controls group, then choose the type of control you want to add (e.g., buttons, checkboxes).
- Assign macros to these controls to make them interactive. For example, you can create a macro that filters data based on user selections.

6.Design Your Dashboard:

Arrange your PivotTables, charts, slicers, and interactive elements on a single worksheet to create your dashboard layout.

Use text boxes, shapes, and colors to enhance the visual appeal and usability of your dashboard.

7.Testing and Refinement:

Test your dashboard to ensure all interactive elements are working correctly.

Make any necessary refinements to improve functionality and usability.

8. Protect Your Dashboard (Optional):

Once you're satisfied with your dashboard, you can protect it to prevent accidental changes. Go to the "Review" tab and click on "Protect Sheet" or "Protect Workbook."