

Lecture Note #11: Charts Part #1

BUSI 201: Business Data Analysis

Topic 1. Charts and Graphs

Faced with a dataset with multiple variables and thousands of entries, it is often incredibly difficult to even grasp the most basic of information such as a time trend or basic correlations. In the previous lectures, we worked with one tool that assisted us in visualizing data; conditional formatting and sparklines. Excel offers a variety of chart types, each with its own “best-use” cases, which can cause some complications for beginners.

To choose the best type of chart for visualizing data, there are a multitude of factors to consider. Does your data track a variable over time? How is the data measured? Do you plan to plot multiple variables on a single plot? Who is your target audience? Is there a variable or data point that you want to draw attention towards? The choice of chart types and colors will depend on your answers to these (and some more) questions.

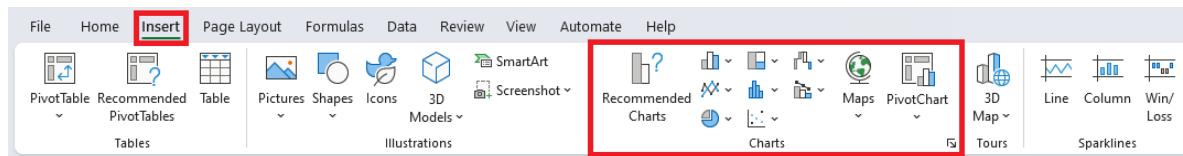


Figure 1: Inserting Charts

You can see a list of some charts that Excel offers by navigating to the **Insert** tab, and observing the elements in the **red box**. In this chapter, we will be examining the four main archetypes of charts, which include...

- Bar Charts
- Line Charts
- Scatter Plots
- Pie Charts

Topic 2. Bar Charts

Bar charts will plot numerical values as a “bar” for each category determined by the data as displayed in Figure 2. One axis will hold information on the category, and the other will hold information on the numerical value. Each category will have its own bar, and the length of said bar will depend on the magnitude of the numerical value. For instance, Figure 2 is a bar graph generated using public use data from the Illinois Department of Transportation.

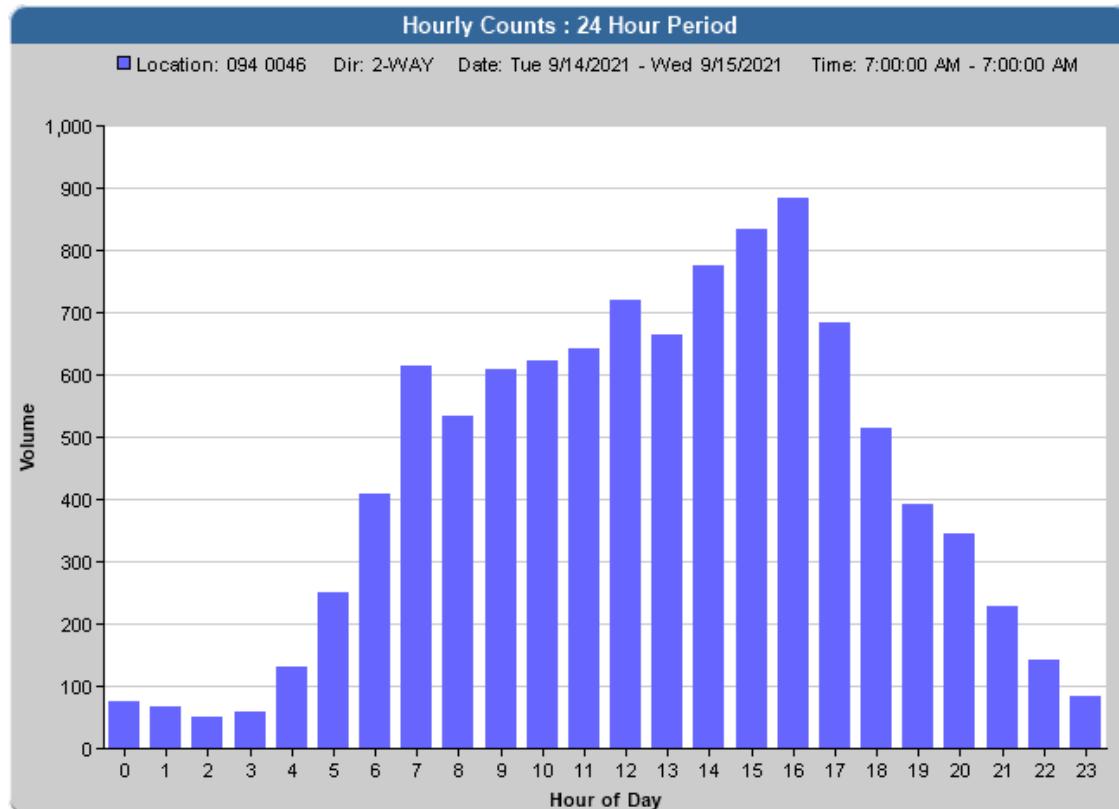


Figure 2: Bar Chart: Traffic Data by Time of Day (IDOT)

To contrast the bar chart, the raw data used to create this plot is organized in the following table.

| 0:00 | 1:00 | 2:00 | 3:00 | 4:00 | 5:00 | 6:00 | 7:00 | 8:00 | 9:00 | 10:00 | 11:00 |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 76 | 67 | 49 | 58 | 131 | 251 | 409 | 614 | 534 | 608 | 621 | 643 |
| 12:00 | 13:00 | 14:00 | 15:00 | 16:00 | 17:00 | 18:00 | 19:00 | 20:00 | 21:00 | 22:00 | 23:00 |
| 720 | 665 | 776 | 832 | 883 | 682 | 515 | 391 | 344 | 229 | 143 | 82 |

Bar Charts: When To Use

Bar charts are most often used when you would like to show the distribution of the data as our example in Figure 2, or when we want compare the level difference between subgroups in your data. When there are subgroups in your data, and you want to easily compare the numerical values assigned to each subgroup, the bar chart is a valid option.

In the previous example, each hour of the day name is the “subgroup,” and their respective traffic count is the “numerical value.” By plotting the data as a bar chart, we can easily see which time of the day is the busiest, when is the least busy period, how many cars were spotted at each hour, etc.

Please navigate to the BAR worksheet in the workbook BUSI201-LEC12-Workbook.xlsx. This worksheet contains sales target data for 15 employees in some firm. We will add a default styled bar graph by first selecting the data (including the variable names) as shown in Figure 3, navigating to the Insert tab in the red box, and clicking the bar chart button in the blue box. Next, choose the basic Clustered Column format in the orange box, and Excel will automatically create a bar chart shown in the green box.

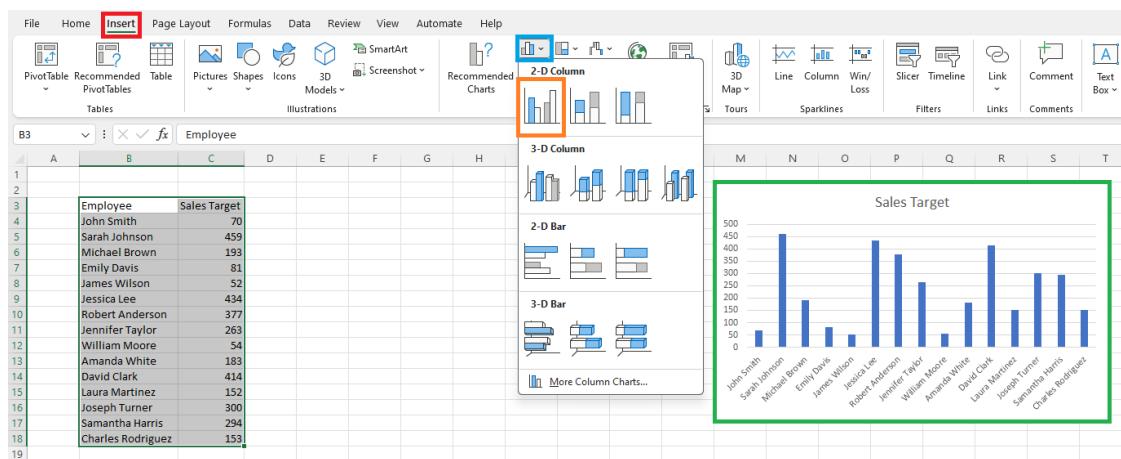


Figure 3: Bar Chart: Sales Target by Employee Name

In this new bar graph that we created, note that the “subgroups” are the employees, and the numerical values are their respective sales targets. For any bar chart, understanding what your subgroups are, and what the numerical values are will be critical, as they are the two axes of the chart itself.

Bar Charts: Data Structure

As a rule of thumb, you should always first identify what your “subgroups” are going to be before plotting a bar chart. The subgroups will almost always be a categorical variable. Some examples of categorical variables may be individuals’ names, departments, months, quarters, industry type, states, etc. Even when a continuous variable is used as the “subgroups,” we will always be transforming them to categorical variables, as shown in the example of traffic data where we categorized the time of the day into discrete “hours.” The counterpart to our subgroups must be numerical data, as this value will determine the length of each bar.

Bar Charts: Elements

Click on the chart that we just created in worksheet BAR, and there will be three buttons that pop up next to the chart as shown in Figure 4. When creating a chart, Excel will automatically choose some elements that it thinks should be useful in visualizing data. We can see that for this bar graph, Excel pre-populated the chart with Axes, Chart Title, and Gridlines.

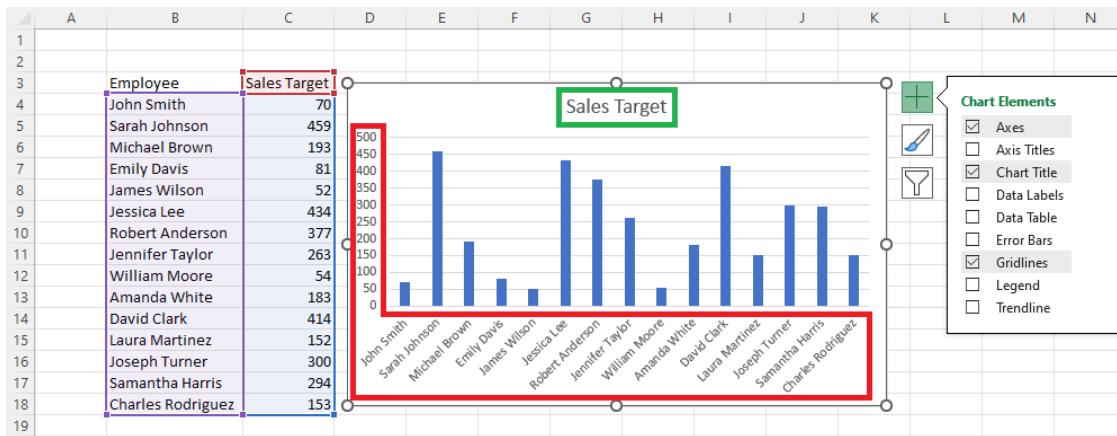


Figure 4: Bar Chart: Elements

A. Axes

Here, the Axes are in the red box, and would be recommended for almost every single bar chart. It clearly communicates the identity of the subgroups and the scale of the numeric variables, which is essential information to communicate.

B. Chart Title

The Chart Title in the green box is a bit more situational. Suppose you intend to export this chart to a Powerpoint presentation. If you have the title of a slide already populated with "Sales Target," having the title appear again in the chart will be redundant information, and would not be required.

C. Gridlines

Gridlines assist in "extending" the axis indicator values with grey lines so that it is easier to accurately measure the length (or height) of each bar. For bar graphs with many subgroups, it may be slightly difficult to eyeball the values for bars on the far side of the axis numbers. Therefore, maintaining the grey gridlines may be useful in certain situations.

D. Axis Titles

Axis Titles are also situational in their importance. If we take a look at our current chart, the horizontal axis is self-explanatory, since it lists the names of the employees. Click on Axis Titles to add titles to each axis, and double-click the added axis titles to add subtitles that provide more information on the units of sales target. Is it measured in dollars? number of units? new clients?

E. Data Labels

A nice alternative to Gridlines is Data Labels. Data labels in bar charts adds a display of the numerical value for each subgroup on top of (or sometimes inside of) each bar. Therefore, we can add data labels and remove gridlines, and click on the vertical axis and **delete** to edit the chart.

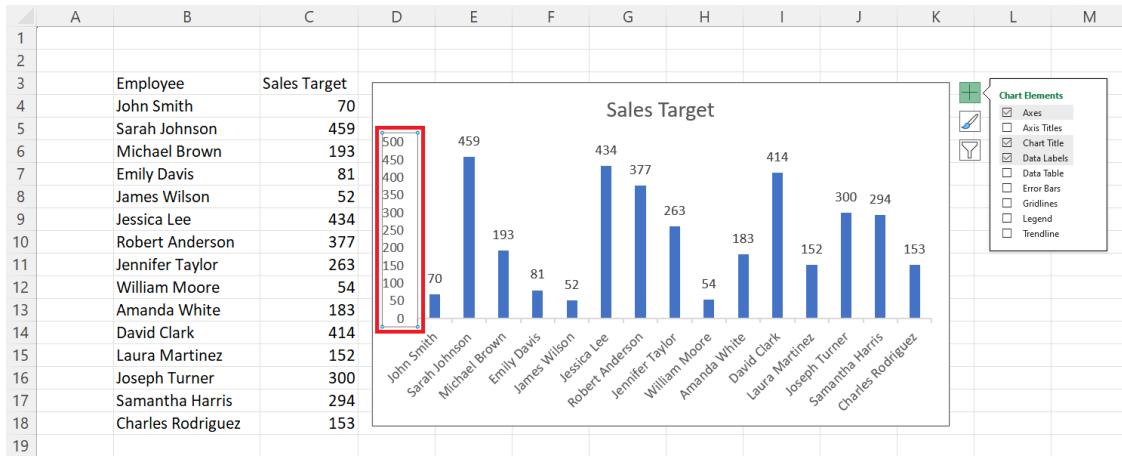


Figure 5: Bar Chart: Labels not Gridlines

F. Legend

For a bar chart that plots a single variable, such as the case of our example, the legend is not necessary. Since there is only one variable, and the **Chart Title** or **Axis Titles** can fulfill the role of legends. However, when we are plotting more than one variable in a single chart, a legend may be necessary. We will return to the topic of chart legends later once we talk about other types of bar charts.

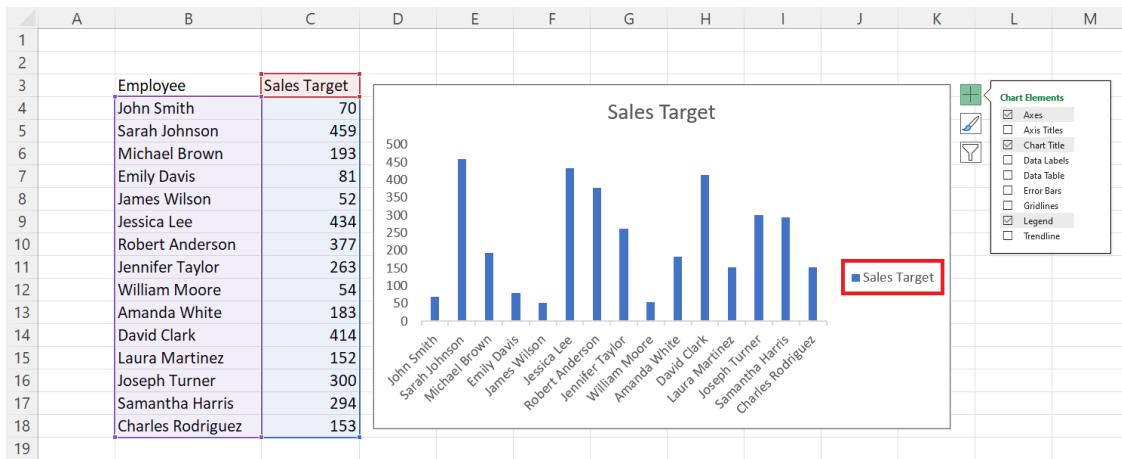


Figure 6: Bar Chart: Legend

G. Data Tables, Error Bars

Data Tables will overlay/add a table with the original data into the existing bar chart. There are not many use cases for Data Tables, as it is preferable to present the visualization and raw data separately. Meanwhile, error bars add a measure of variability to the end of each bar when the numeric values are summary measures (e.g. averages), but not frequently used as there are charts better suited for these purposes.

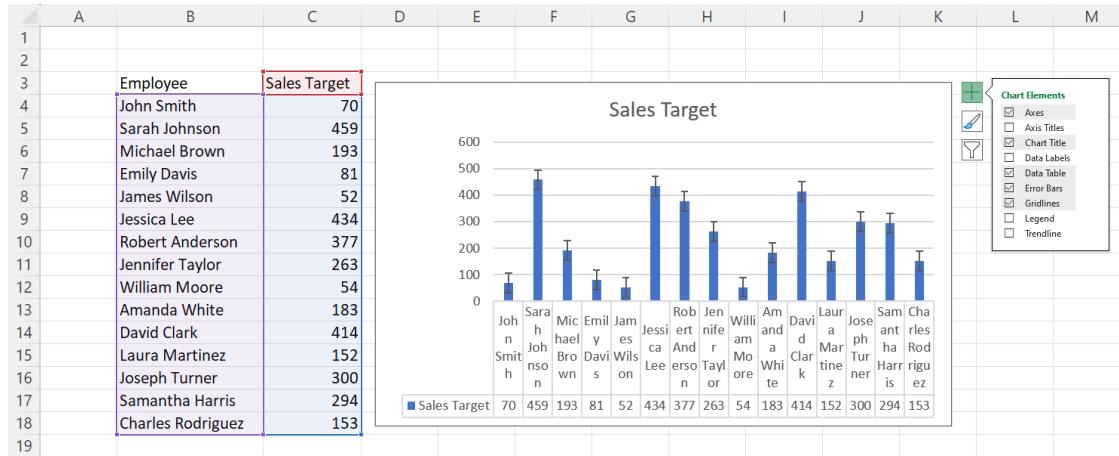


Figure 7: Bar Chart: Data Tables and Error Bars

H. Trendlines

Trendlines will assume that there is a trend starting from the left-most element, that sequentially leads to the right-most element. This may be helpful if the subcategories are dates or years, where we can assume that there is a time trend. However, in our example of sales targets by employees, adding a trend is nonsensical.

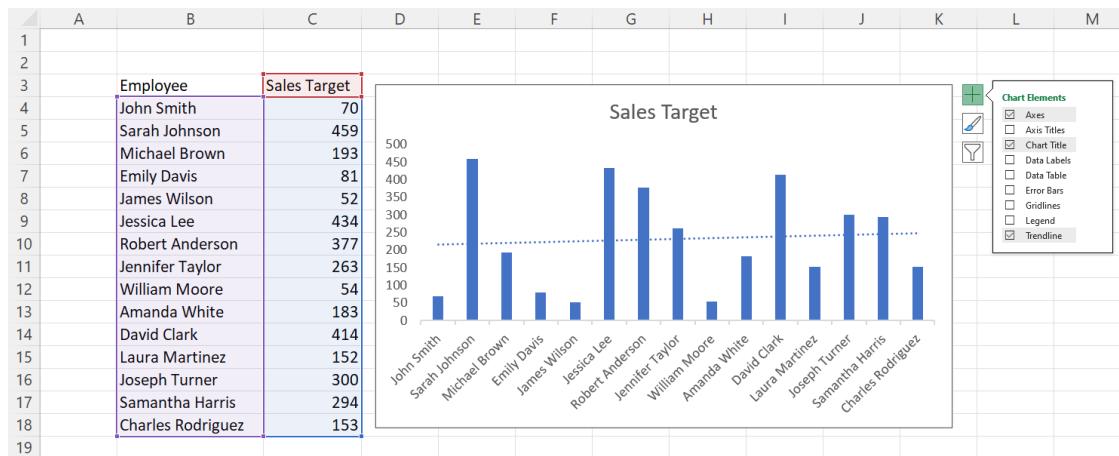


Figure 8: Bar Chart: Trendlines

Bar Charts: Customizing

As is most other things Excel, you are not stuck with the default format that is given to you when you first create a chart. We will briefly discuss some of the changes that one can make to the “canned” charts. Click on the chart, and navigate to the Chart Design tab that appears.

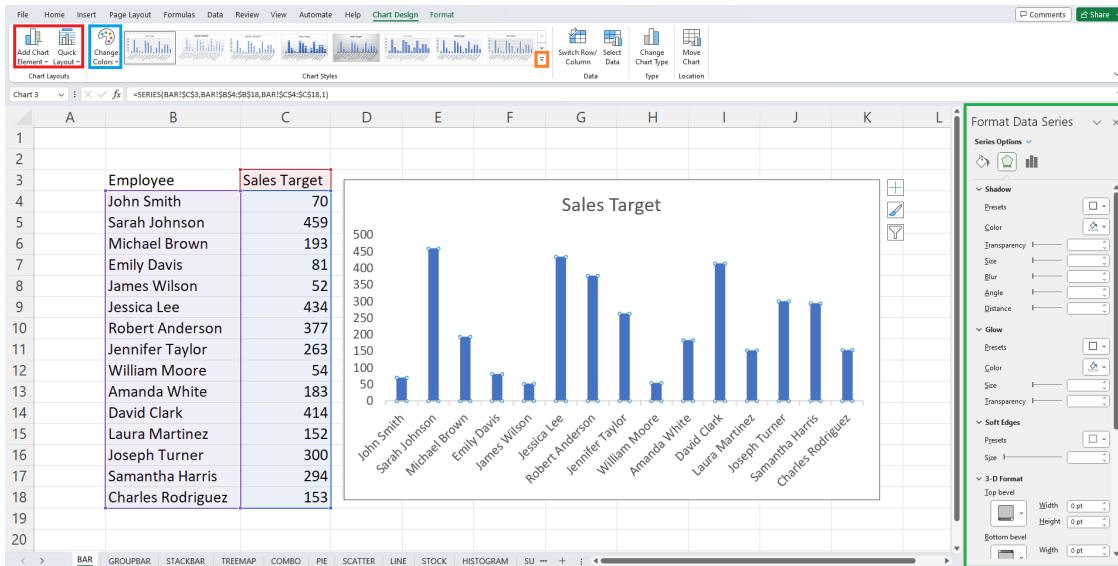


Figure 9: Bar Chart: Chart Design

In the red box are the options to add and eliminate elements from the chart. Clicking the blue box allows us to change the color scheme of the bars to some preset color palettes. Expanding the menu by clicking the orange box will give the user some options to change the theme of the entire chart. If these options are not to your liking, you can always manually change each element. You can try out customizing the chart by working with the tools in the green box. See Figure 10 for an example.

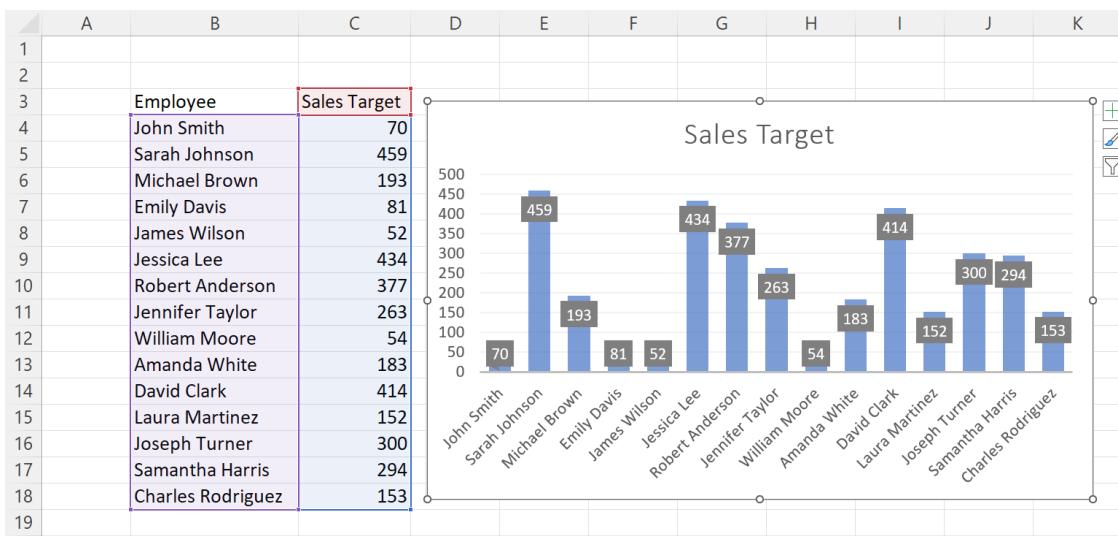


Figure 10: Bar Chart: Custom Chart Design

Bar Charts: Sorting Data?

Another thing that deserve consideration whether to sort the data in one way or another before creating the chart. Compare the two charts in Figure 11. Sorting the data by numeric value before creating the chart takes the burden off from the viewer, which may be a desirable trait.

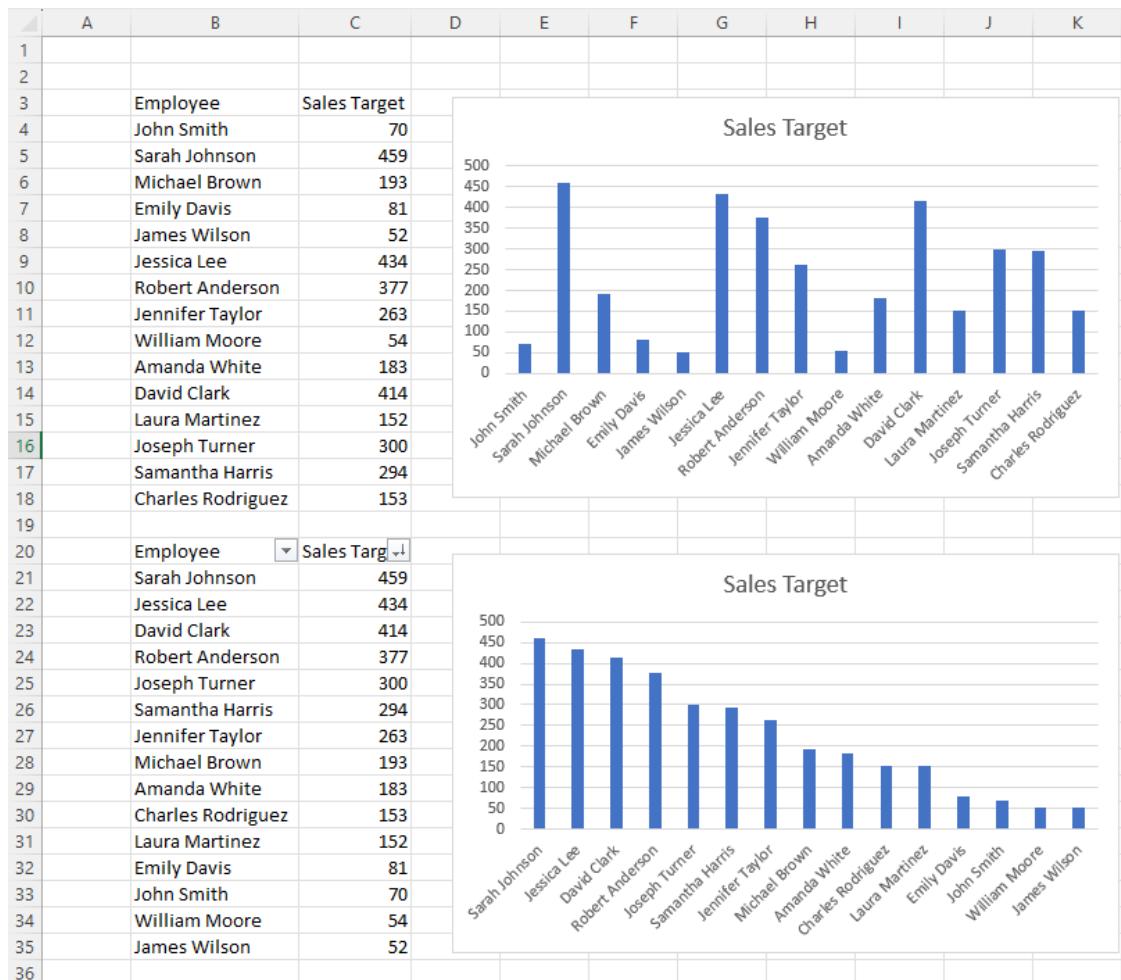


Figure 11: Bar Chart: Sorted vs. Unsorted Data

However, we must also be careful when sorting data based on the numeric value. Suppose that we sorted the traffic data from Figure 2 by numeric value. It would make absolutely no sense, as the sub-groups we were using were “hour of day,” and having the hours in a proper time series is more important than sorting based on the observed number of vehicles.

As is true with most aspects of data visualization and presentation, these decisions are highly contextual to the data, target audience, and the message you wish to convey.

Topic 3. Bar Charts: An Extension

Previously, we examined how bar charts work, when to use bar charts, and how to format bar charts. In this topic, we will be examining two different types of bar charts, one where we have more than one numerical value for each subgroup, and another where we break down each bar into distinct parts.

Multiple Numerical Values

Navigate to the worksheet GROUPBAR, where there is the same table that populated the previous worksheet, but with another variable appended. Suppose that we want to plot both the employees' sales target, and their actual sales for the year in a single chart.

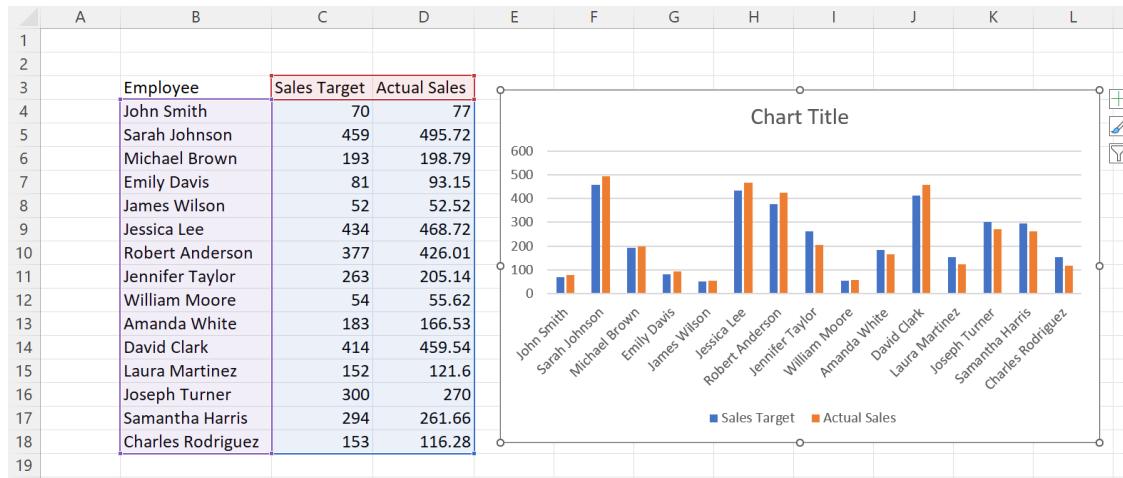


Figure 12: Bar Chart: 2 Numerical Values

You can have more than 2 numerical values using the data in the same worksheet. For instance, Figure 13 shows a bar chart with three numerical values per each subgroup. While it is possible to add even more variables for each subgroup, we should be careful.

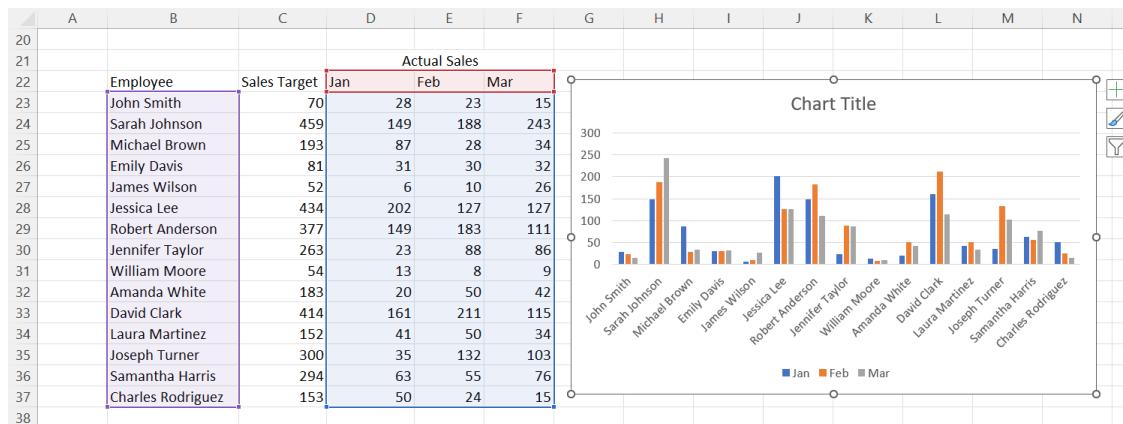


Figure 13: Bar Chart: 3 Numerical Values

If you overdo adding numerical variables to each subgroup, you may end up with a chart that is completely not legible. See Figure 14 for an example of why adding too many numerical variables might backfire. Here we have 12 numerical variables for each individual, representing the realized sales information for each month of the year.

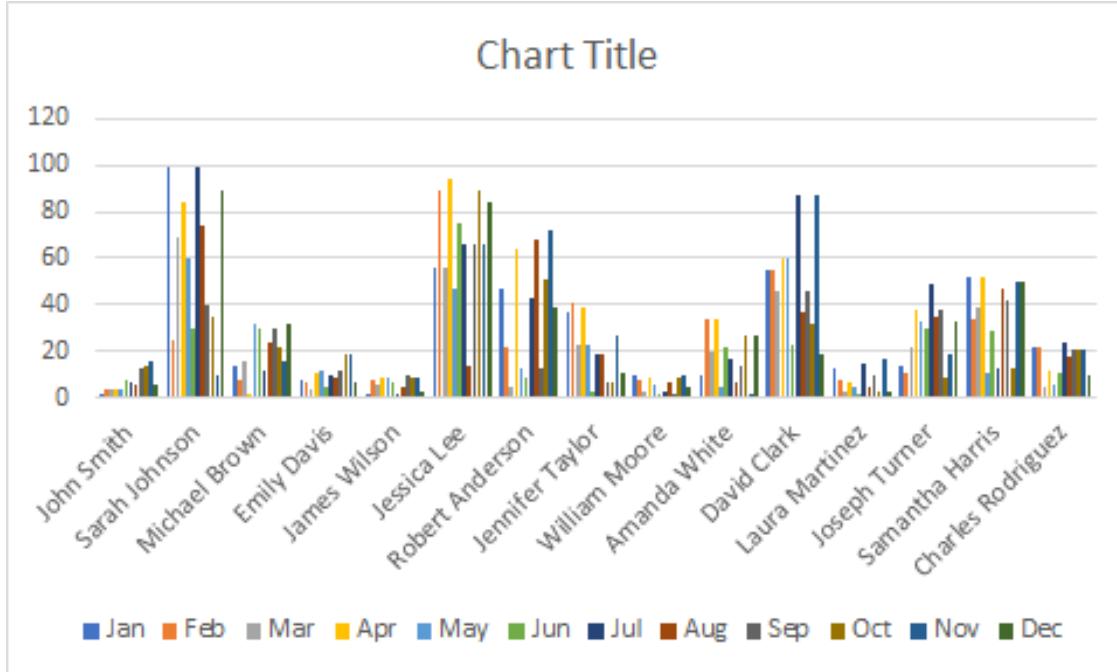


Figure 14: Bar Chart: 12 Numerical Values

An alternative way of visualizing this data may be to aggregate information into longer “blocks.” For instance, in Figure 15, we aggregate the 12 months into 4 quarters, and have four numerical variables for each subgroup. While the chart remains quite busy, it is comparatively easy to read compared to Figure 14.

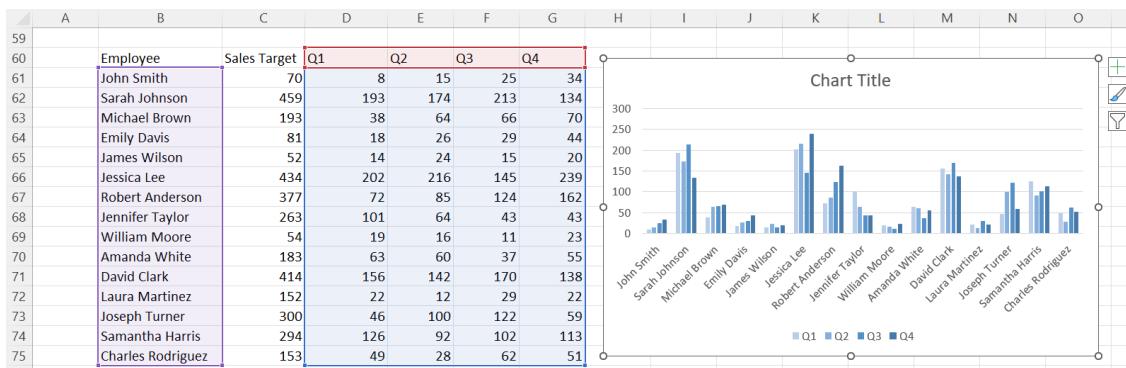


Figure 15: Bar Chart: 12 Numerical Values

Stacked Bars

The second case that we will be examining is useful when the main numerical value that you wish to communicate can be broken down to smaller parts. For instance, suppose that you have the data in worksheet STACKBAR. This data breaks down the employee's first quarter sales target by months. If you simply plot the first quarter target as we do in Figure 16, we lose some vital information about the monthly goals.

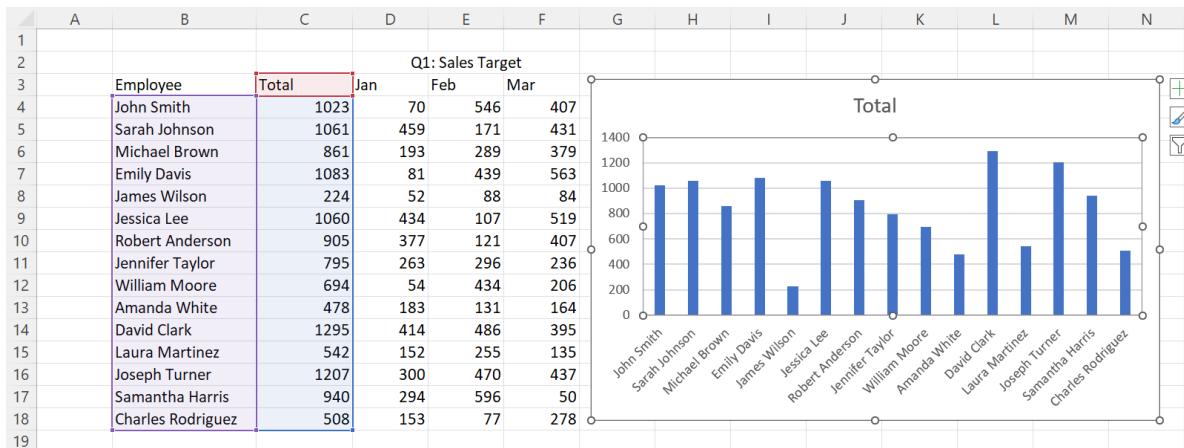


Figure 16: Bar Chart: 12 Numerical Values

Meanwhile, if we use the method introduced previously to add bars based on numerical values side-by-side as in Figure 17, we can see each employee's monthly target. However, now it is not easy to spot the differences in the total 1st quarter sales targets for each employee.

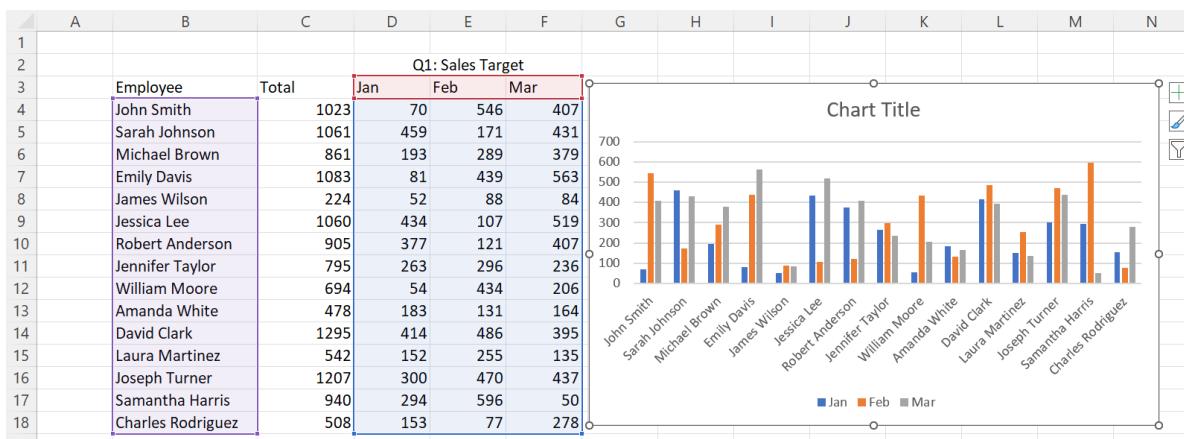


Figure 17: Bar Chart: 12 Numerical Values

In situations like this, we can stack the numerical values vertically, and have them show up like Figure 18. We can clearly see both the total sales target for each employee, and also see the breakdown of each employee's monthly targets.

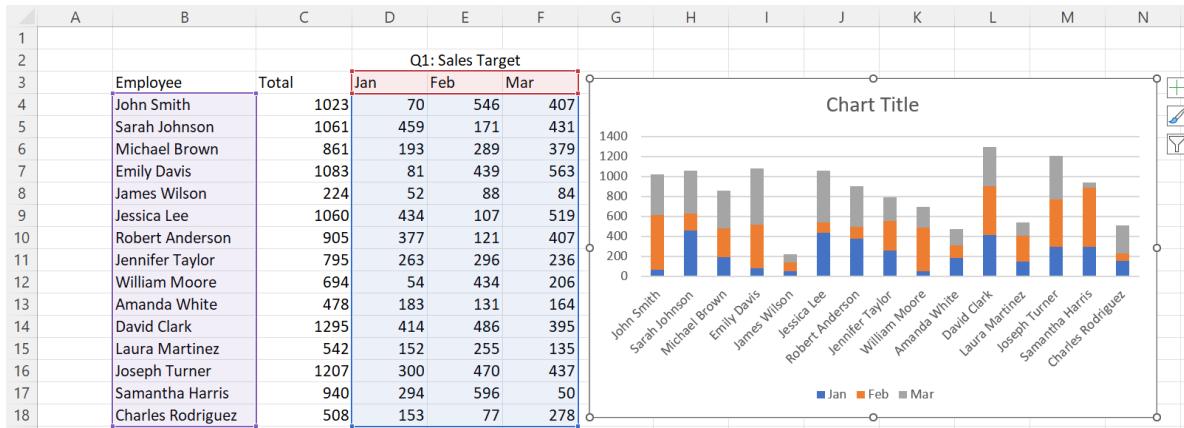


Figure 18: Bar Chart: Vertically Stacked

To plot the “vertically stacked” bar graph, see Figure 19. Select the column that contains information on the employee’s names, and the monthly sales targets. To select these non-contiguous zones, first select any of the two zones first, and select the remaining zone while pressing down the **ctrl** key. Once you selected the correct zone, then navigate to **Insert**, and select the bar chart in the **red box**. Then, you should have the stacked chart show up.

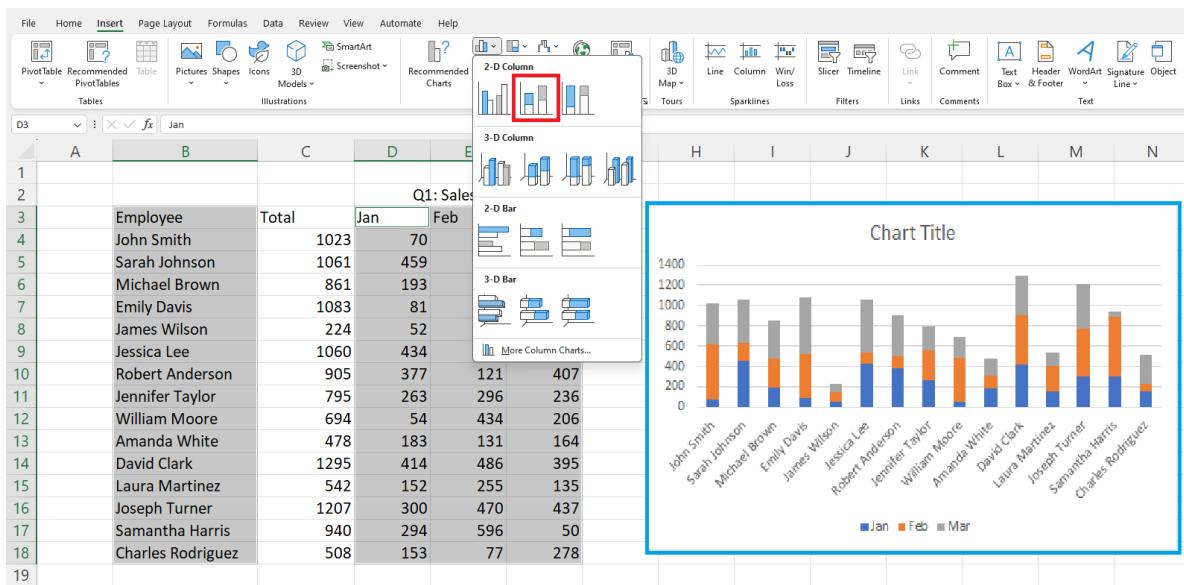


Figure 19: Bar Chart: Vertically Stacked

Editing and Removing Variables

If you have trouble selecting two non-contiguous zones using the **[ctrl]** key, you may try out the following steps. First, simply select the entire dataset, and create the stacked bar chart. Notice that in the **green box**, that the first quarter total sales is included in the plot.

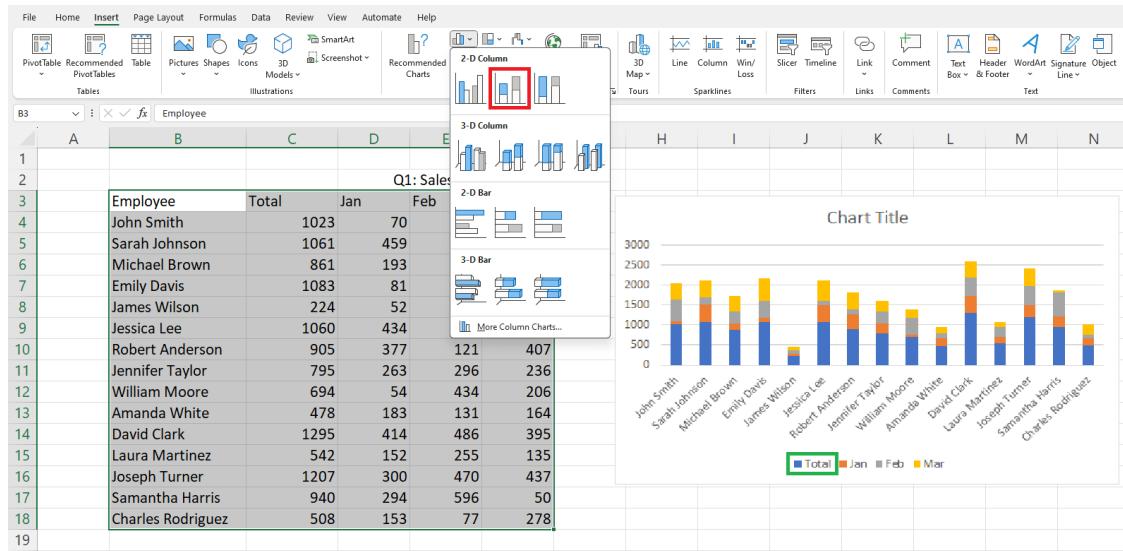


Figure 20: Bar Chart: Incorrect Chart

To remove this Total from our chart, select, and right click on the chart and click Select Data in the **red box**. Then, in the pop-up window named Select Data Source, select the variable to remove, Total in the **orange box**, and click Remove in the **green box**. Select OK, and then Total will be removed from the chart.h

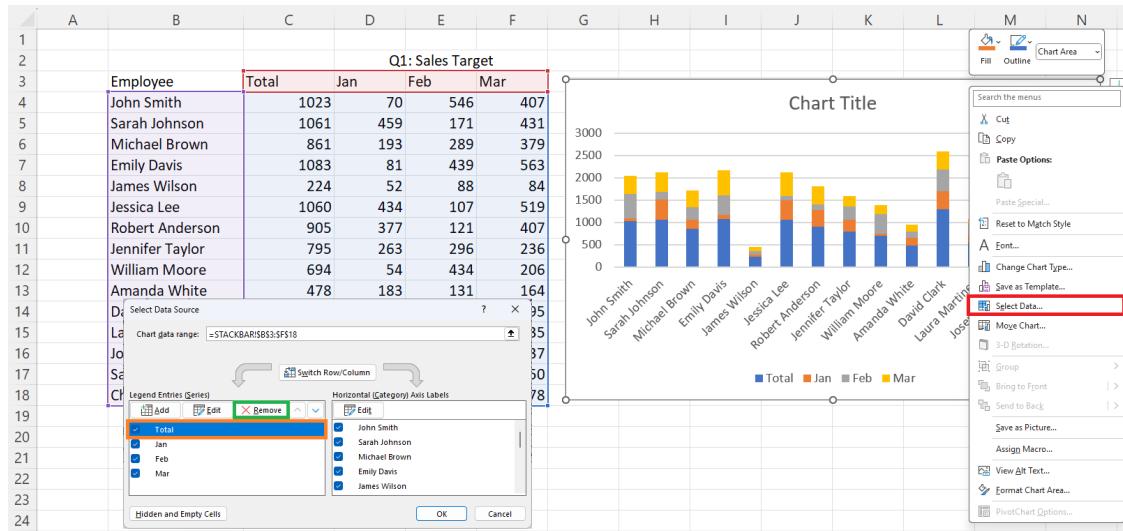


Figure 21: Bar Chart: Editing the Incorrect Chart

100% Stacked Bars

Navigate to the worksheet ENERGY, which contains data on electricity generation in the US from 1950 to 2010 by decades, based on the source of energy. Applying the previous stacked bar chart, we have Figure 22. This graph shows the viewer how the total energy generated in the US evolved over time, and how the composition of each energy source shifted over time as well.

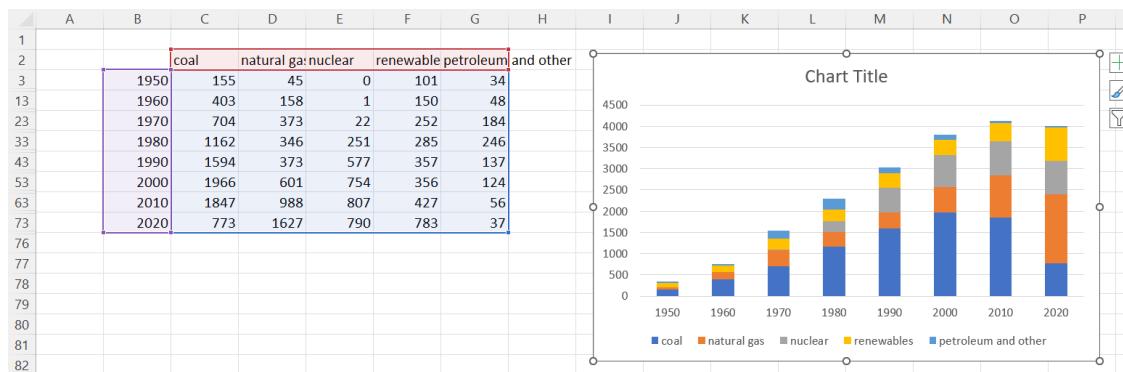


Figure 22: Bar Chart: Energy Generation by Source / Decade

However, if you want to concentrate on the share of each source, while ignoring the changes in total energy generation, you may want to use a different type of graph, the 100% Stacked Bar chart. The 100% stacked bar chart will automatically calculate the “share” of each number for each subgroup, and create a bar chart that has the same length (height), and display the proportion of each number.

You can plot the 100% stacked bar chart by following the steps in Figure 23, and the resulting chart is also visible in the same figure. Note that while you gain some clarity on the changing composition of the energy sources by adopting the 100% stacked bar chart, you do lose out on some critical information about total energy generation.

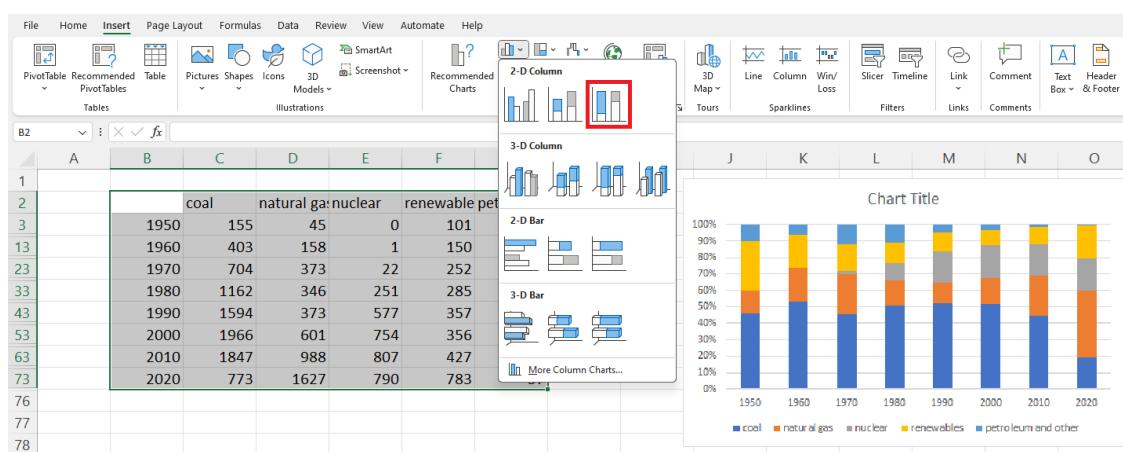


Figure 23: Bar Chart: 100% Stacked Bar Chart