

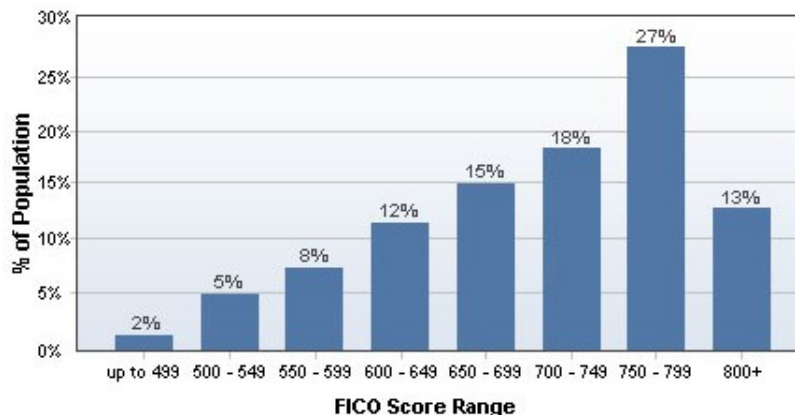
LECTURE 07: OF DATA AND DISPLAYS

I. Quantitative Data

- a. Both kinds of data displays we'll discuss here makes use of "bins." A bin is a range of numbers with a certain number of observations falling into that range.
- b. In all cases of bins, you can't tell what's going on inside of it. If 1,000 salaries are between \$30,000 and \$60,000, you don't know if most of those salaries are closer to \$60,000, closer to \$30,000 or spread out evenly or some other distribution entirely.
- c. Bins are useful because they can turn any number of observations into a manageable range that you can use for graphing.

II. Histogram

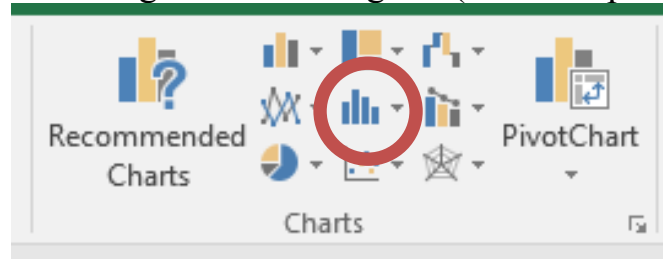
- a. *Histogram*—a histogram divides data into groups and displays the number of observations per group
 - i. **The number or percent of observations in each bin is not constant**—all bins have a number or percent of observations based on the bin range. If two bins have the same number or percent of observations, it's a coincidence.
 - ii. **The range of each bin is constant**—each range of a bin has the same value, such as \$10,000 or 50 points.
 - iii. Advantage: Easily organizes lots of data, especially when there are many possible divisions (e.g. income or other continuous variable)



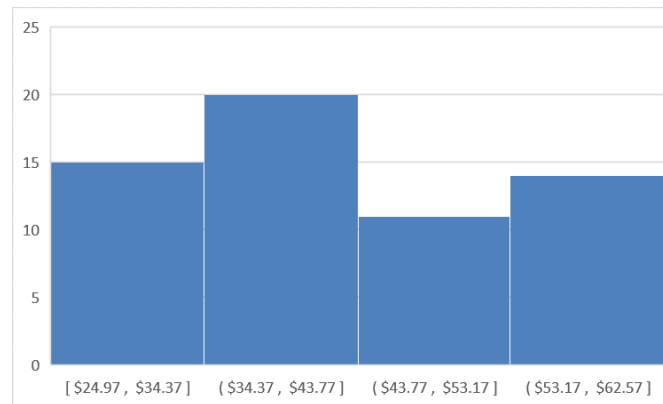
III. Making Histograms

- a. There's a simple way to create a histogram in Excel but you have to jiggle with it a bit to make it look nice.

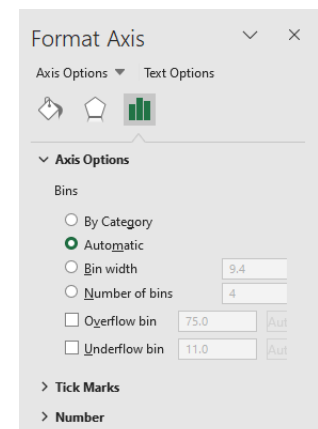
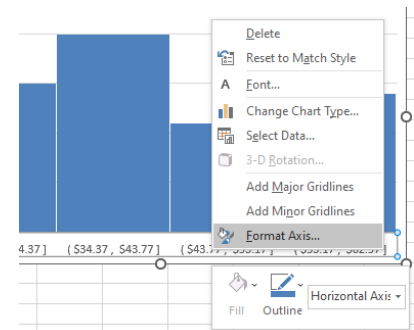
- b. To create a histogram, highlight the data you want to use and go to Insert >> Histogram >> Histogram (the first option).



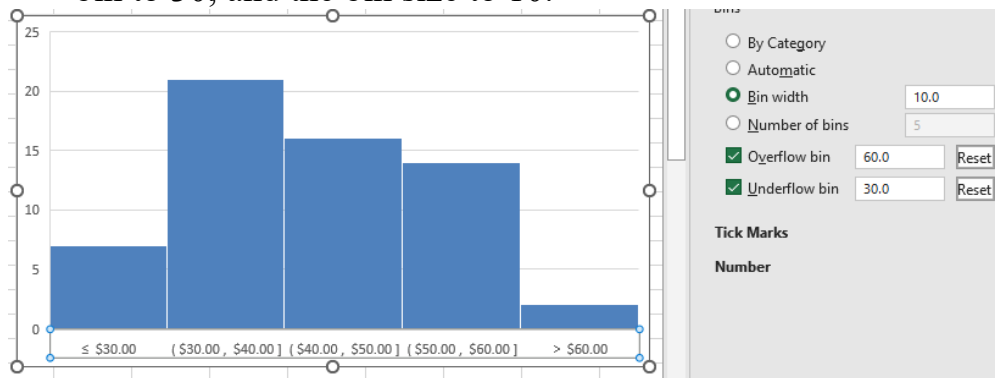
- c. For example, here's a histogram of Delta's monthly stock price.



- i. Note: brackets are inclusive and parentheses are exclusive. Thus \$34.37 would be in the farthest left column.
- d. Notice the problem: the ranges are awkwardly defined. While they correctly have a uniform interval (\$9.40 in this case), they are directly based on the raw data and thus look strange.
- e. To fix this, click the bins, right click, then select Format Axis...
- f. This'll open up a panel on the right, allowing you to force the bin range (called Bin Width) to a certain level. You can also change the number of bins and even add "overflow" and "underflow" bins.
- i. The overflow bin lets you set a number and put all observations above that number in it.



- ii. The underflow bin does the same as the overflow but it's all observations at or below that number.
- g. Note that “bin width” and “number of bins” uses the radio button, which means you can't have them both active at the same time. If you set the number of bins, Excel will force the bin range to a certain value; this is what I mean by you having to jigger with it a little to get the result you want.
- h. I strongly recommend using both the overflow and underflow bins to get the desired bins. Here, I set the overflow bin to 60, the underflow bin to 30, and the bin size to 10.

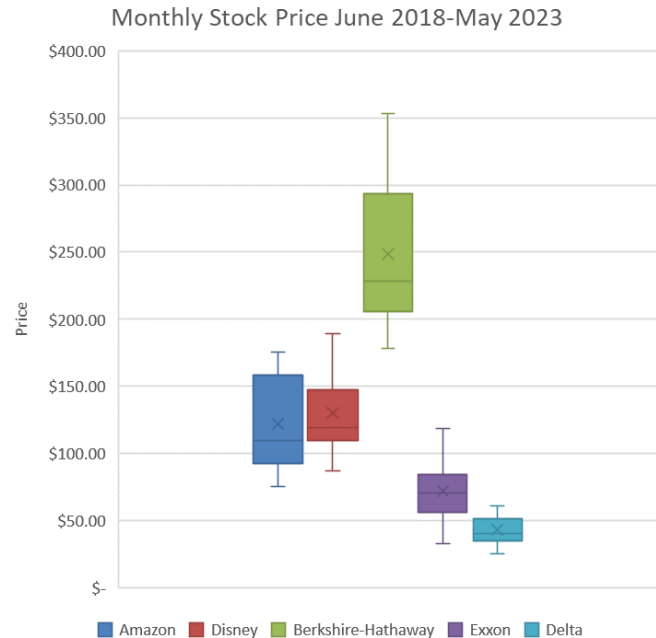


IV. Box Plot

- a. *Box Plot*—a display which shows where quartiles of data are
 - i. **The number or percent of observations in each bin is constant**—all bins have one-fourth of the total observations (a quartile).
 - ii. **The range of each bin is not constant**—bin range must vary to accommodate the distribution of the data, thus you can infer the distribution of the data from the bin range. If two bins have the same range, it's a coincidence.
 - iii. The 1st quartile is a data value which indicates where, from the minimum to that value, are the first fourth of the observations.
 - iv. The lines on either side of the box show the range between the maximum and 3rd quartile and between the minimum and 1st quartile.
 - v. The box is between the 1st and 3rd quartile with a line (the median, or 2nd quartile); the box is called the *interquartile range*.
 - vi. Advantage: It illustrates dispersion but it is able to handle virtually any number of observations. All you need to make a box plot are five numbers: maximum, minimum, 1st quartile, 3rd quartile, and median (2nd quartile).

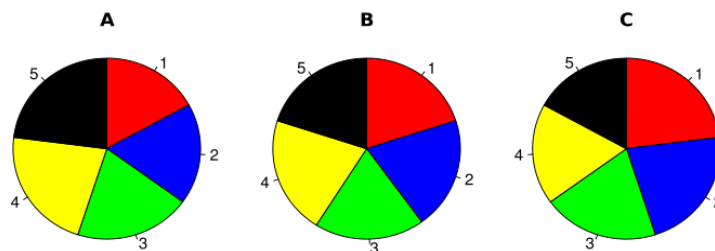
V. Making A Box Plot

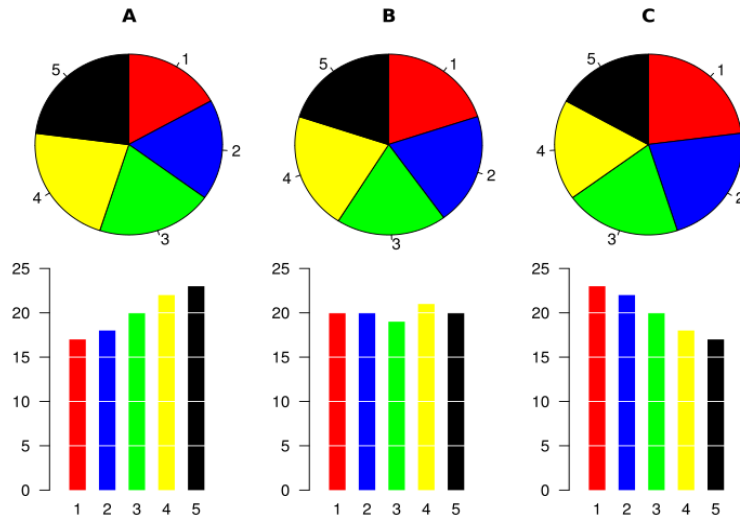
- a. To create a box plot, highlight the data you want to use and go to Insert >> Histogram >> Box and Whisker (the second option).
- b. Box plots are usually displayed horizontally but for some reason, Excel makes them vertical. (The “X” indicates the average.)



VI. Categorical Data

- a. All of these previous types of displays help us organize data given as a continuous variable, such as a number. But sometimes you want to organize *categorical data*, where there are several groups and the data consists of how many observations are in each group.
- b. *Pie Chart*—a circular chart divided into sections, or wedges, describing a percent of total each group is. Bigger wedges mean a bigger percent. This is one of the most widely used charts out there but it's not perfect (as I will show you).
 - iii. Advantage: It is widely used and easy to understand.
- c. *Bar Chart*—like a histogram, but each bar represents a category rather than a range of a distribution (in a way, each distribution is a category).
 - i. Advantage: It is also widely used and easy to understand. It typically has an advantage over bar charts in showing each group's size relative to the other.
- d. In B, is black or green larger?

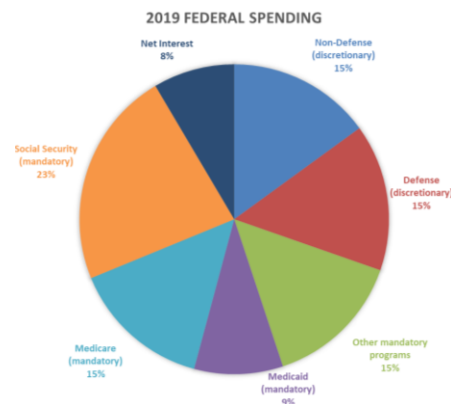
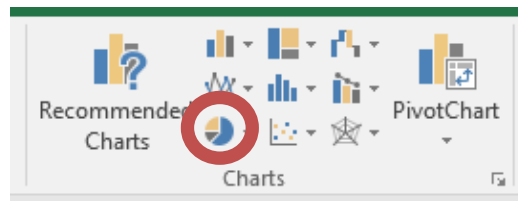




- ii. It's harder to tell in a pie chart that what's going on because in a pie chart, you're comparing two areas while in a bar chart, you're comparing two lines.

VII. Making Pie and Bar Charts

- Recall these charts reflect categorical data. We need some group to put observations in. Use **Data Set 2** for this part.
- For a pie chart, highlight the data for 2019 federal spending and labels and select Insert >> Pie Chart image >> 2D Pie.

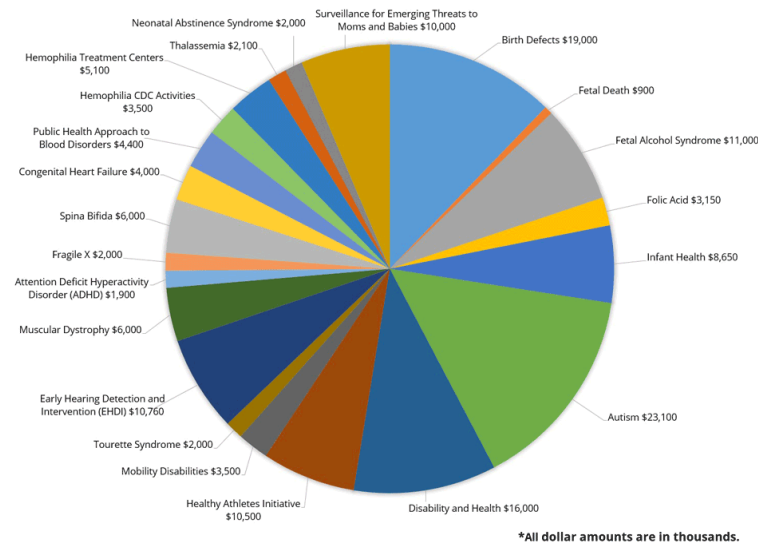


- Note I selected a different style than the default and had to jiggle with the label placement so they didn't overlap with the wedge.

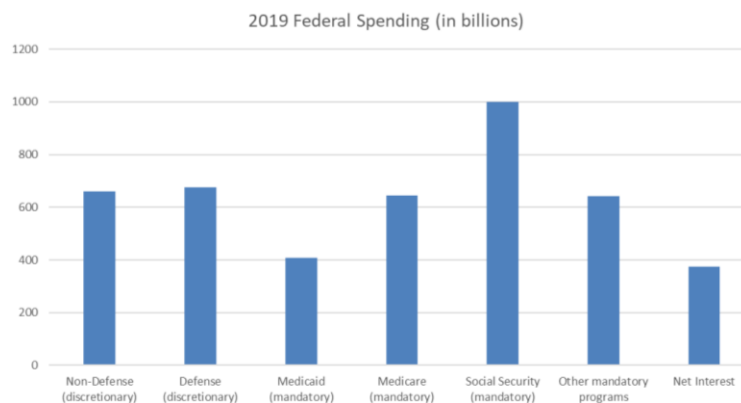
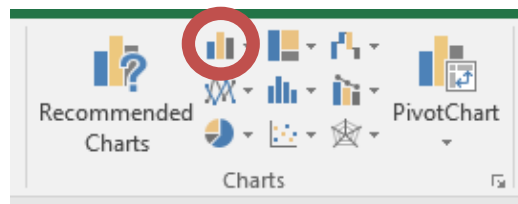
Also note that Excel automatically converted our numbers into percents.

- ii. A word of caution: don't have too many categories. Check out the CDC's 2019 budget:

<https://www.cdc.gov/ncbddd/aboutus/budget/index.html>

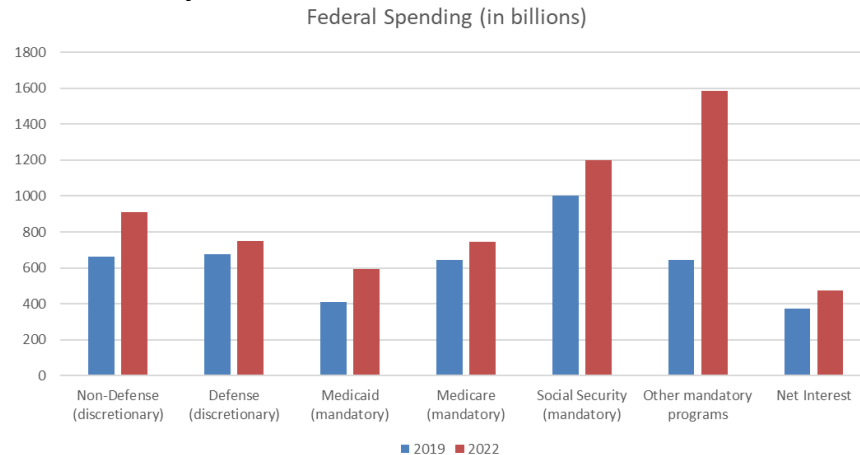


- c. For a bar chart, highlight the data of 2019 and labels and select Insert Column or Bar Chart >> Clustered Column.



- i. The line chart is a bit bulky but, as before, it's easier to tell which section is larger.

- ii. Plus, it can have multiple years. Highlight the labels and data for both years:

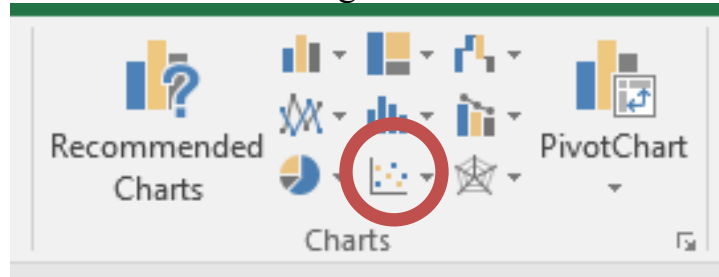


- iii. Note the huge increase in “other mandatory programs,” which included \$482 billion for the Biden Administration’s attempt to forgive student loans; it was added to the budget in September 2022.

VIII. Scatterplot

- [Here's](#) a video tutorial of making a scatterplot.
- A scatterplot has two axes, each with a variable; each dot represents an observation, sometimes labeled with the observation’s element.
 - The variables should be adjusted for population, when appropriate. To determine if it’s appropriate to adjust for population, ask yourself if the variable should change simply because there are more people. Total number of births and total number of crimes should be adjusted for population. Rainfall and percent literate should not be.
 - To adjust for population, divide the gross number (gross meaning total, not disgusting) by the observations’ population.
 - If the resulting values are very small, multiply each observation by some appropriate number, such as 1,000 or 100,000. This will give you the number of (say) births per 1,000 people or per 100,000 people.
- Scatterplots help us see how variables are (or are not) related. Consider this website, which lets you manipulate an elaborate scatterplot: [https://www.gapminder.org/tools/#\\$chart-type=bubbles&url=v1](https://www.gapminder.org/tools/#$chart-type=bubbles&url=v1)
- You’ll notice on Gapminder that you can express a variable on a linear (lin) or logarithmic (log) scale.
 - A linear scale means each unit is some previous unit plus a fixed value. For example: 10; 20; 30; 40; 50; etc.

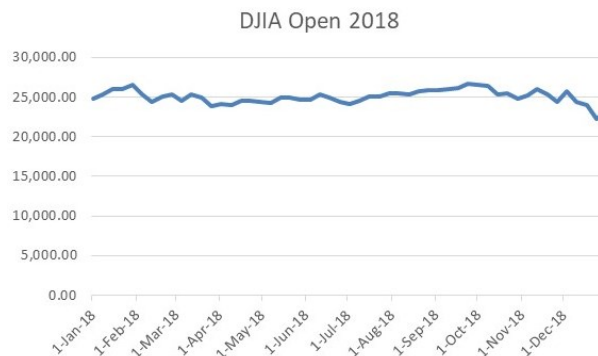
- ii. A logarithmic scale means each unit is some previous unit *times* a fixed value. For example: 10; 100; 1,000; 10,000; etc
- iii. For values with a wide range, where the largest value might be orders of magnitude larger than the smallest value logarithmic scales are a better visual choice.
- e. To make a scatterplot in Excel, go to the country data tab in **Data Set 2** and highlight columns G and H (murder rate and pop density).
- f. Then Insert >> Scatter image >> Scatter.



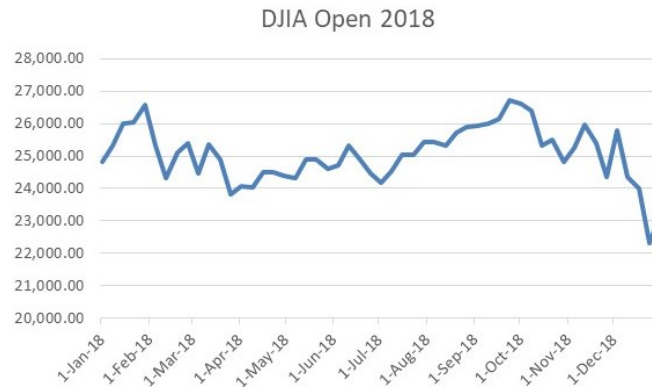
- i. Excel defaults whatever variable was on the right as the vertical axis and title. Whatever's on the left is the horizontal axis.
- g. You'll notice that while some observations stand out, we can't really tell what's going on. We need to transform population density using natural log.
- h. Excel makes this easy. Click the population density axis and then right click it. Select Format Axis. You'll see a logarithmic option appear on the right side of the screen. Click it to turn it on.

IX. Truncating Axes

- a. The range of the axes on charts can be changed, usually done by truncating, or cutting off, part of the y axis. A truncated graph's y axis does not start at zero; this enables easier reading of the graph.
- b. For example, considering this line graph of the opening weekly values of the Dow Jones Industrial Average for the year of 2018.



- c. It's hard to see how much the values are changing over the years. Let's change it by changing the y axis.
 - i. To do this, select anywhere on the y axis and right click, selecting Format Axis.
 - ii. Under Bounds, let's change the minimum to 20,000.



- iii. Now we can see what's going on week-to-week.
- d. Excel defaults by truncating the y axis, though truncation comes with dangers. While the above diagram is more readable, the DJIA looks more volatile than it is. The lesson is that you should always watch the y axis for truncation. Deceptive truncation is one of the ways people lie with statistics.
- e. Another example: labor force participation rate for [men](#) and [women](#).