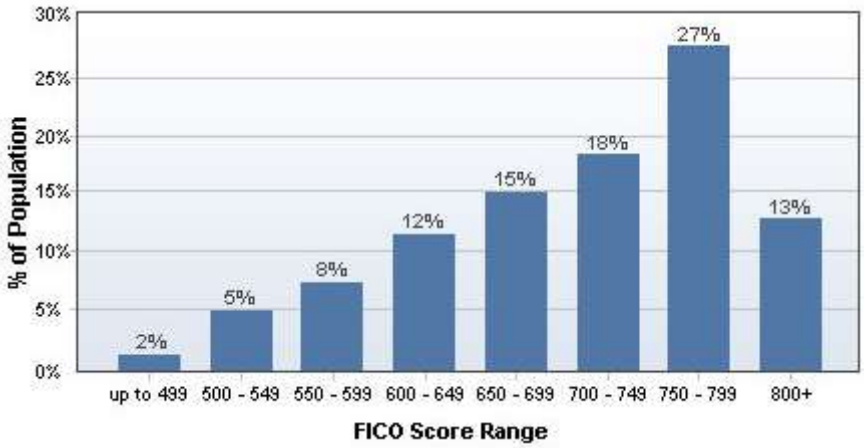
David Youngberg
BSAD 210—Montgomery College

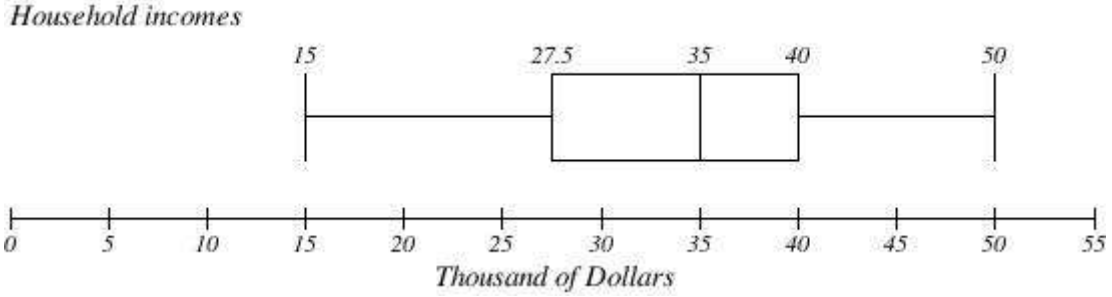
LECTURE 07: OF DATA AND DISPLAYS

I. Abbreviations
   a. Before we get into data displays, one theme that will come up is the trade-off between information and simplicity. A display which conveys a lot of information is good because it allows people to learn more. But it's also bad because it can get confusing.
   b. One way to make displays more approachable is to cut down on the number of numerals is uses. A graph with a maximum number of 30,000,000,000 is harder to read than a graph with a maximum number of 30.
      i. All those zeros not only take up a lot of space, the reader has to count the commas to figure out if it's thirty million or thirty billion or thirty trillion.
   c. As such, displays (and tables) will often abbreviate values with phrases like "in thousands" or "in millions." I'll also often ask homework answers to be in thousands or in millions, too. So let's be clear what that means.
   d. Imagine the "in thousands", etc. represents the appropriate comma, replacing the decimal point. Thus:
      i. 5,000 in thousands is 5
      ii. 7,531,800 in thousands is 7,531.8
      iii. 7,531,800 in millions is 7.5318
      iv. 98,050,000 in millions is 98.05
   e. There are different practices on how to represent "in thousands", etc. with a single letter. The standard I'll use for this class is:
      i. K = thousands (50K = 50,000)
      ii. M = millions (118.1M = 118,100,000)
      iii. T = trillions (76.56T = 76,560,000,000)

II. Quantitative Data
   a. *Histogram*—a histogram divides data into groups and displays the number of observations per group
      i. Advantage: Easily organizes lots of data, especially when there are many possible divisions (e.g. income or other continuous variable)



b. *Box Plot*—a display which shows where quartiles of data are
   i. A quartile is a part of a data set with one-fourth of the total observations. The 1st quartile is a data value which indicates where, from the minimum to that value, are the first fourth of the observations are
   ii. Note you can also divide the data into other segments such as in five equal parts (quintiles), ten equal parts (deciles), one hundred equal parts (percentiles), etc.
   iii. The lines on either side of the box show the range between the maximum and 3rd quartile and between the minimum and 1st quartile
   iv. The box is between the 1st and 3rd quartile with a line (the median, or 2nd quartile); the box is the *interquartile range*.
   v. The larger the distance between these points, the more disperse the observations. The shorter the distance, the more concentrated
   vi. <u>Advantage:</u> 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).



III. 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.

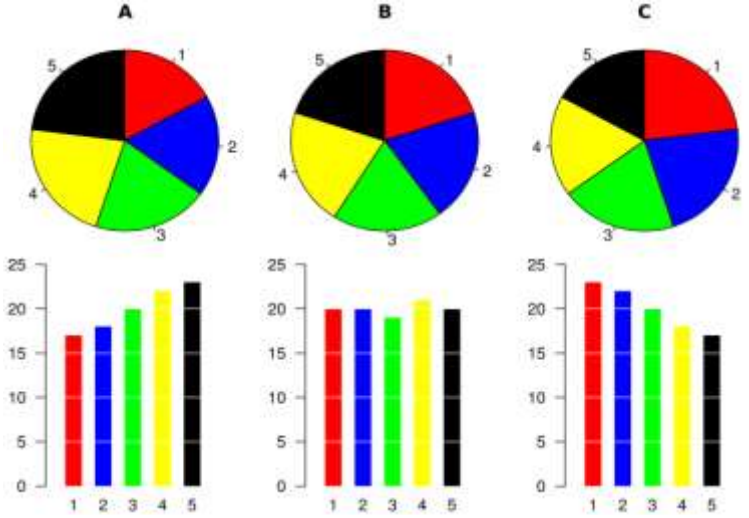
b. *Pie Chart*—a circular chart divided into sections, or wedges, describing a percent of total each group is. Bigger wedges mean a bigger percent. This is one of the most widely used charts out there but it's not perfect (as I will show you).

   i. 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 red or blue larger?



   i. It's harder to tell in a pie chart that they are equal sizes.

IV. Scatterplot

a. The first step in any research project is finding data (this sometimes occurs even before you know what you want to investigate).

b. The second step is determining your approach to the data.

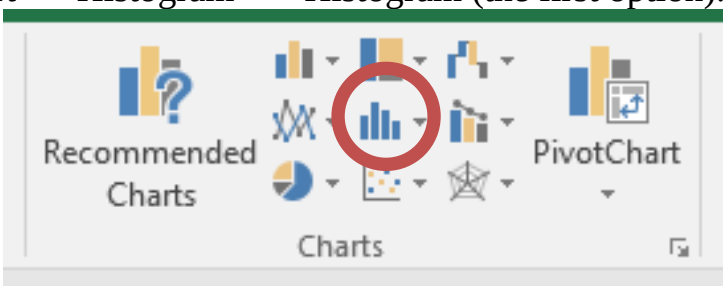
c. A *scatter diagram* indicates how two (or more, if you are feeling adventurous) values relate to each other.

d. Gapminder (www.gapminder.org) is an excellent resource to explore relations between different variables. The website employs data from all over the world to various sophisticated scatter plots. The raw data are available in Excel format.

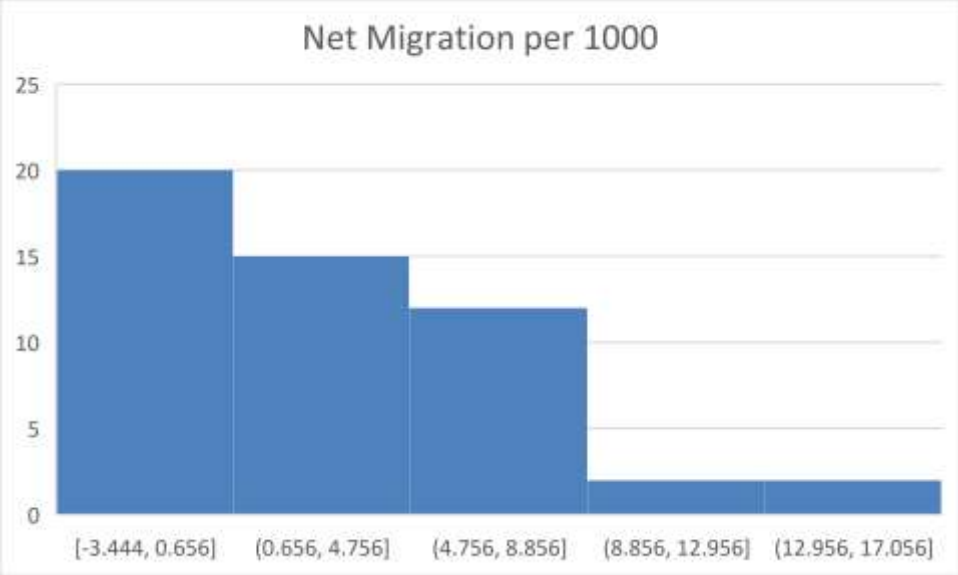
e. You'll notice on Gapminder that you can express a variable on a linear (lin) or logarithmic (log) scale.
    i. A linear scale means each unit is some previous unit plus a fixed value. For example: 10; 20; 30; 40; 50; etc.
    ii. A logarithmic scale means each unit is some previous unit *times* a fixed value. For example: 10; 100; 1,000; 10,000; etc
    iii. For values with a wide range (especially ones skewed right) logarithmic scales are a better visual choice.

V. Creating displays Practice
a. Open Data Set 1, found on my website. This is cross-sectional data of 184 countries and seven variables.
    i. Keep in mind the descriptions tab at the bottom of the page if you want to know more about what each variable is.
b. I use ">" to indicate click order. For example, Page Layout > Margins > Normal means click Page Layout, then Margins, then Normal.
c. It's always a good idea to add labels. You can find how to add labels (notably the horizontal label, the vertical label, and the title) in the formatting area after you make a display.

VI. Histograms
a. There's a simple way and an advanced way to create a histogram in Excel. We will focus on the simple way, however the advanced way is helpful because, as you'll see, the simple way doesn't create a nice-looking diagram.
    i. If you want to know how to fix this, I'm happy to tell you but it's rather time-consuming so I'm leaving it out. But you should be aware of the quick way's limitations.
b. To create a histogram, highlight the data you want to use and go to Insert >> Histogram >> Histogram (the first option).



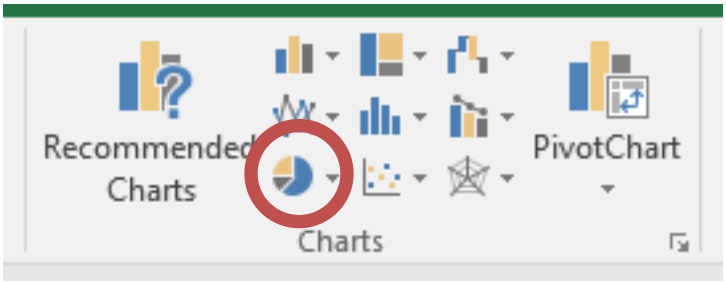
c. For example, here's a histogram of 2012 net migration per 1,000 people, based on U.S. states (and the District of Colombia).

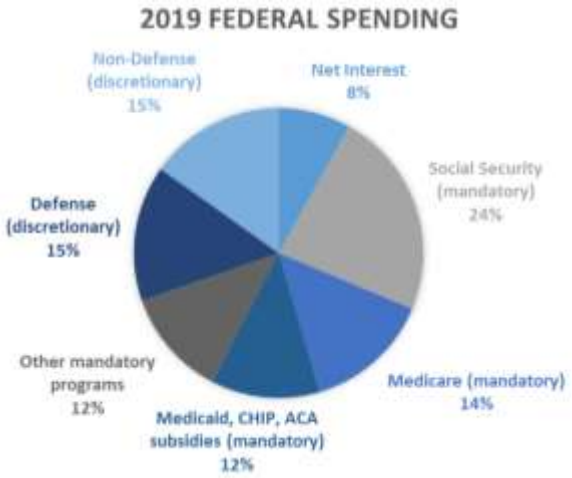


d. Notice the problem: the ranges are awkwardly defined. While they correctly have a uniform interval (4.1), they are directly based on the raw data and thus look strange. And there's no way to directly fix this.
e. To fix this, you have to create your own ranges, called bins, and use the histogram option in Data Analysis. This takes a while to explain so we're just going to move on but here's a video that explains it if you're curious. (You will not be tested over how to make a histogram the advanced way, but you should know how to activate Data Analysis; see below.)
    i. This technique uses the Data Analysis function, which you can find under the Data tab. If you don't see it, you should learn how to activate it because we will use Data Analysis later in the semester.
    ii. Here's how you do it:
        1. File > Options > Add-Ins > Go…
        2. Click Analysis Toolpak and OK. You might have to install it if you've never activated it before.
    iii. If you don't see Data Analysis (it should be on the far right of the Data tab), turn off the Toolpak and then reactivate it. It should show up then.

VII. Pie Charts

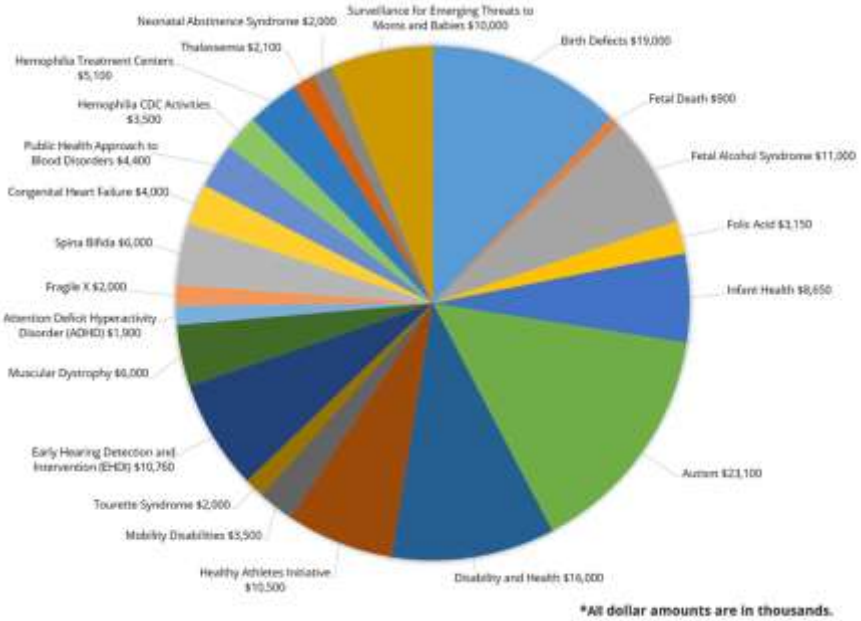
a. Recall these charts reflect categorical data. We need some group to put observations in. Let's use US federal spending for 2019.
b. Highlight the histogram output and select Insert >> Pie Chart image>> 2D Pie.





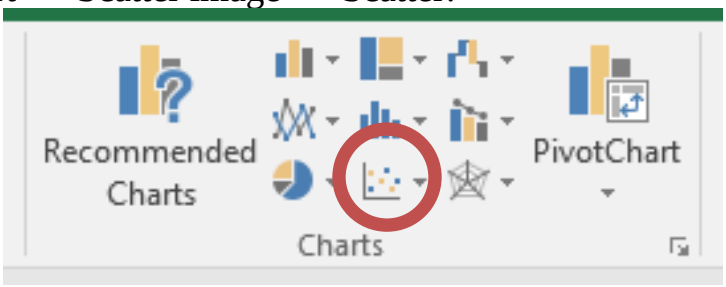
c. Again, make sure your categories column is descriptive.

VIII. Also, don't have too many categories. Check out the CDC's 2019 budget: https://www.cdc.gov/ncbddd/aboutus/budget/index.html



IX. Scatterplot

a. Here's a video tutorial of making a scatterplot.
b. First highlighting columns G and H (murder rate and pop density).
c. Insert >> Scatter image >> Scatter.

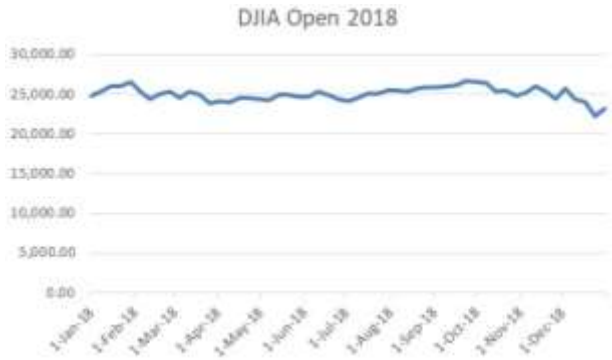


   i. Excel defaults whatever variable was on the right as the vertical axis and title. Whatever's on the left is the horizontal axis.

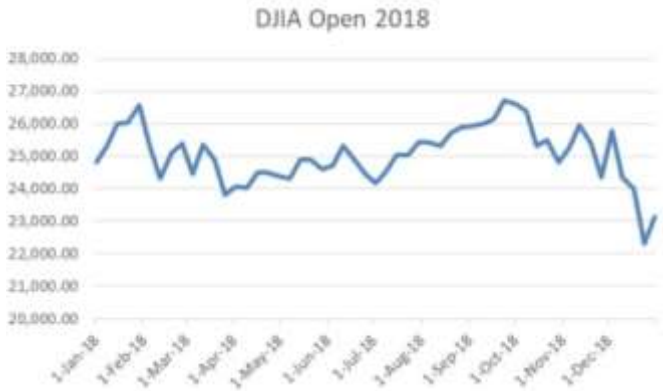
d. You'll notice that while some observations stand out, we can't really tell what's going on. We need to transform population density using natural log.
e. Excel makes this easy. Click the population density axis and then right click it. Select Format Axis. You'll see a logarithmic option appear on the right side of the screen. Click it.

X. 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 by gender.

XI. Printing

a. If you want to print an image, click it and try to print it. Excel will print just the image you've selected.