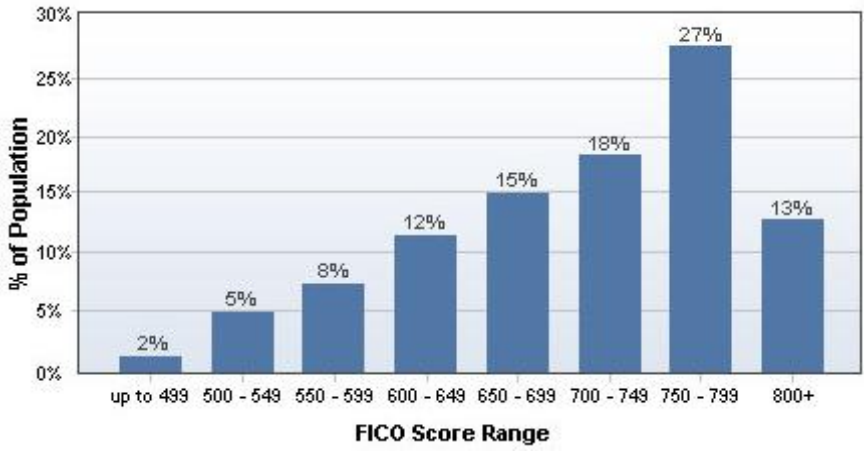
David Youngberg
BSAD 210—Montgomery College

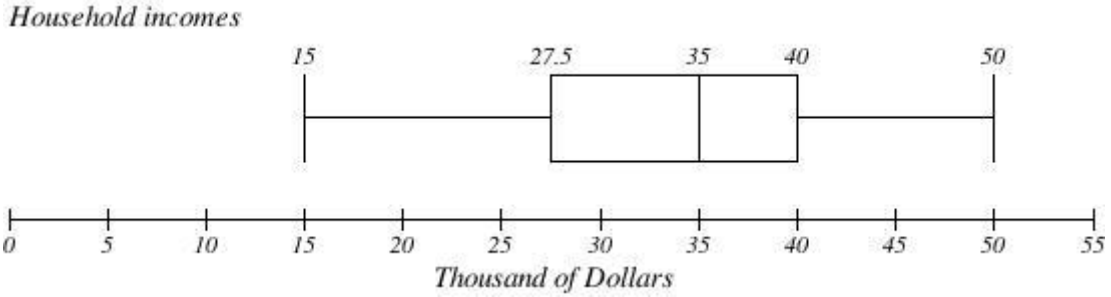
# LECTURE 04: OF DATA AND DISPLAYS

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> Like the steam-and-leaf diagram, 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.

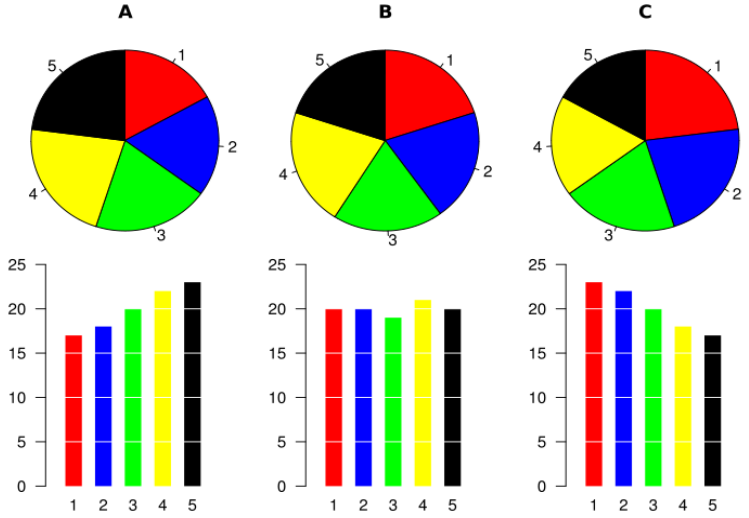
b. *Pie Chart*—a circular chart divided into sections, or wedges, describing a percent of total each group is. Bigger wedges mean a bigger percent. This is one of the most widely used charts out there but it's not perfect (as I will show you).

   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.

   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?

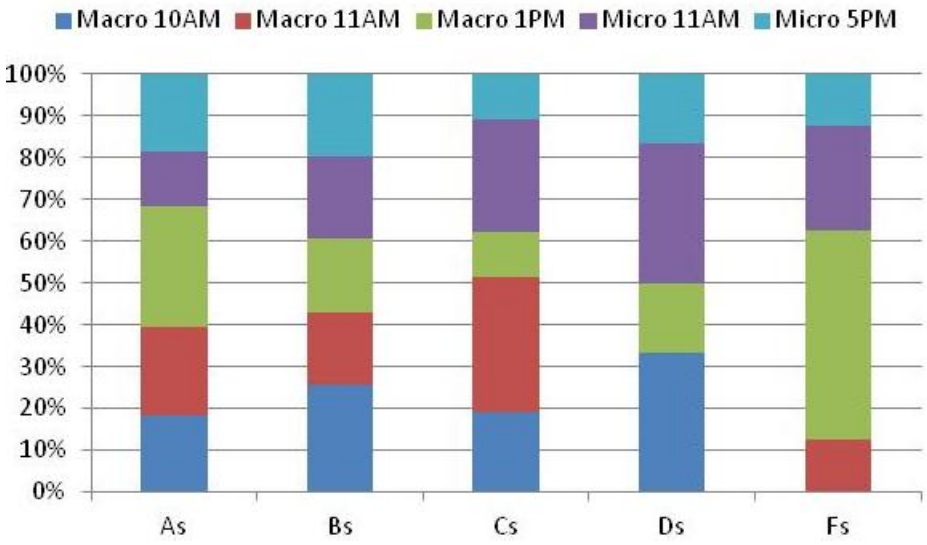



e. *100% Stacked Bar*—this chart sets each group as a bar representing not a value, but 100%. Each group is then divided based on a different category, with the vertical distance determined by the percent.

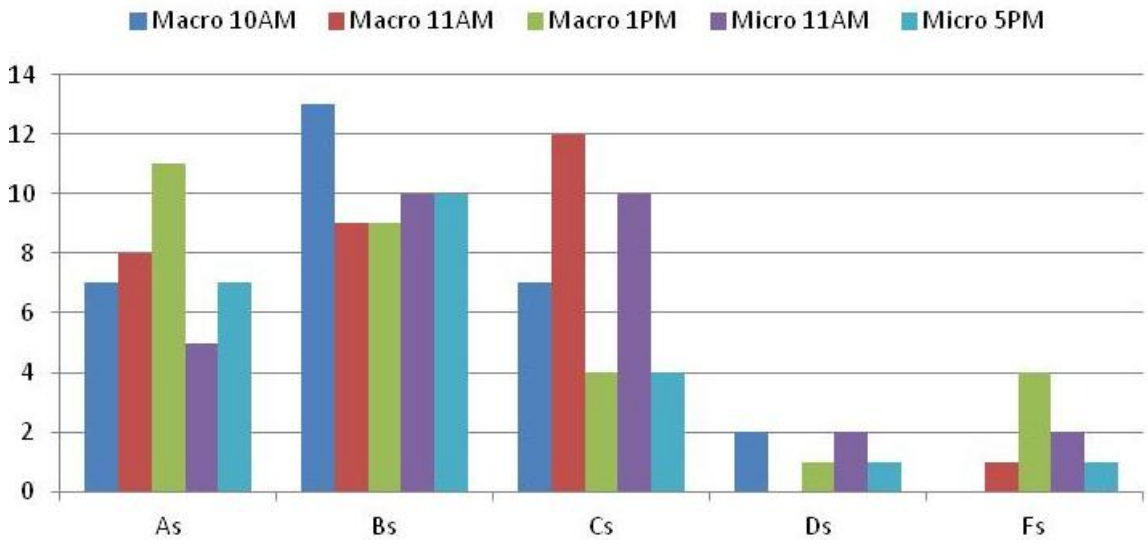
   i. Advantage: It's great for comparing groups in the second-level category and displays in a small amount of space.

   ii. Stacked bars can be a little deceptive, though. Consider this grade data from my classes in the spring of 2014.

f. Which class received the most Fs? Which class did the worst over all?




i. The first question is easy to answer, but the second one is trickier. Consider the bar graph, using the same data:




g. Bar charts are, in general, your best option—it's clear which class got the most Fs and it's also clear Fs overall were unusual—but note that this graph takes a bit more room and looks a bit cluttered. Taking up too much space with too much going on can be a problem, too.

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.

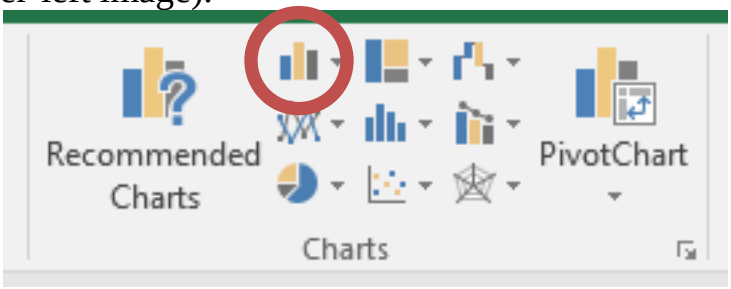
V. Creating displays Practice
a. Open Data Set 1, found on my website. This is cross-sectional data of 184 countries and seven variables.
    i. Keep in mind the descriptions tab at the bottom of the page if you want to know more about what each variable is.
b. Instructions use arrows ">>" to indicate 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. Histogram
a. Here's a short video version of this section.
b. To create a histogram, we need to define our boundaries. Glancing at BMI Male, it appears to range from about 20 to about 30. Let's make each category cover two points: 20 to 22, 22 to 24, etc. Note it's a somewhat arbitrary range; this is one of the weaknesses of a histogram.
    i. In K1, type "Range" and then press ENTER. Then in K2, type 20 and press ENTER. Then, in K3, type 22. Continue this until you reach 32 in K8.
c. Now we need the analysis tool package.
    i. File >> Options >> Add-Ins >> Go…
    ii. Click Analysis Toolpak and OK. You might have to install it if you've never activated it before.
    iii. Data >> Data Analysis >> Histogram

iv. Select column D for the Input Range
v. Highlight K1 to K8
vi. Select Labels to indicate that the first row of the data is the name of the variable.
vii. Select Output Range and select a cell where the information will be printed. I selected K10.
viii. Select OK.

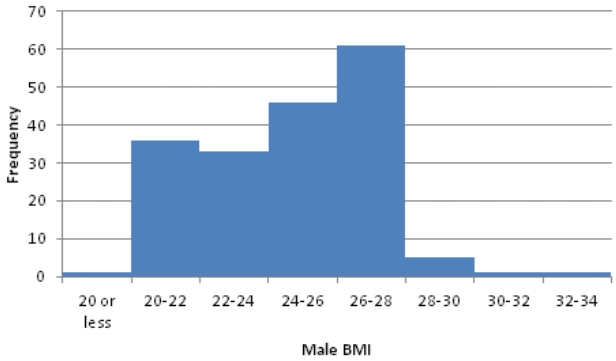
d. Excel has now sorted data into your "bins," or the ranges each bar indicates. We have:
   i. 1 country with a male BMI of 20 or less;
   ii. 36 countries with a BMI greater than 20 but no higher than 22;
   iii. 33 countries with a BMI greater than 22 but no higher than 24;
   iv. 46 countries with a BMI greater than 24 but no higher than 26;
   v. 61 countries with a BMI greater than 26 but no higher than 28;
   vi. 5 countries with a BMI greater than 28 but no higher than 30;
   vii. 1 country with a BMI greater than 30 but no higher than 32;
   viii. 1 country with a BMI greater than 32.

e. Highlight L11 to L18. Then: Insert >> Column image >> Column (upper-left image).




f. But we have terrible labels on the horizontal axis. Let's fix it!
   i. Click a number on the horizontal axis, then right click and select Select Data. Under the Horizontal (Category) Axis Labels, select Edit.
   ii. Highlight K11 to K18 and press OK. Then press OK again.
   iii. Now if you change the values in K11 to K18, the bar graph will change, too.
   iv. You can also add labels under Layout >> Axis Titles.

g. Technically, all histograms have no gap between the bars. Right click any bar and select Format Data Series. Under Series Options, drag Gap Width all the way to 0%. Press ENTER. Here's what I ended up with:

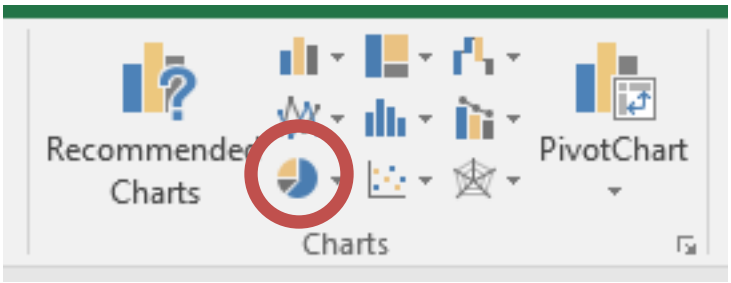


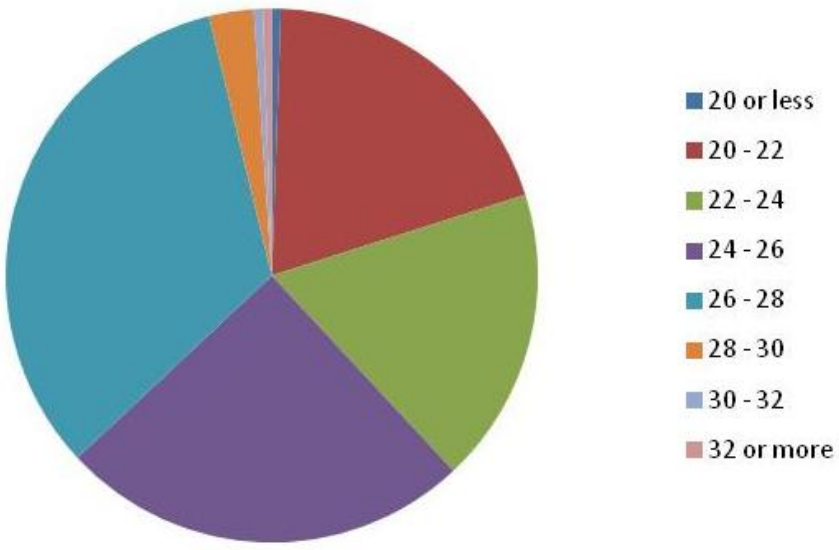

h. Note Excel also has an option to create a histogram with one click. This is not recommended because the bin sizes will be awkwardly set. They will not be easy to read or intuitively spaced.

VII. Pie Charts

a. Recall these charts reflect categorical data. We need some group to put observations in. This data doesn't seem to have anything like that but wait! We just made a category when we were doing histograms!

b. Highlight the histogram output and select Insert >> Pie Chart image>> 2D Pie.

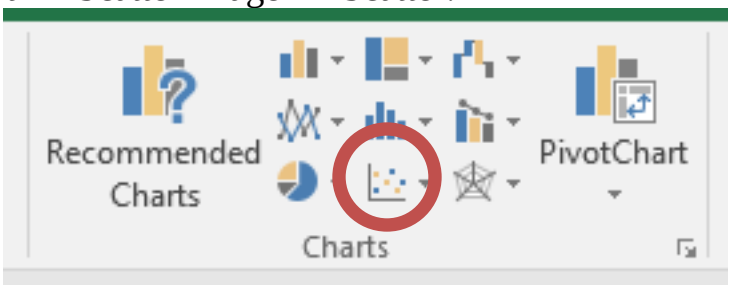


c. Again, make sure your categories column is descriptive. Note the bin values we used to make these categories isn't good enough.

VIII. 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.

IX. 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.