Week 2 Lecture 1:

Displaying and Describing Categorical Data

Review categorical data – a column that describes a category. Pull up example table.

Color, Major, Product, etc.

Frequency tables – counts vs percentages. Excel demo with 311 data. Create counts, then create percentages. This is sometimes called a frequency distribution.

Data visualization –

Why use it? Gives you a sense of what’s going on quicker than simply looking at numbers. If anything interesting jumps out, you know where to dive deeper. Can communicate findings to others much better than tables of numbers and categories.

Bar charts – show example from 311 data displayed earlier. Simple, but often effective. Can be vertical or horizontal (usually base this off size of the labels). Can have values or percentages on the y axis and it doesn’t change the graph. Do we even need numbers most of the time?

First rule of data viz – simplify as much as possible, but no more.

Pie charts – show example. Using our data, then bad example using other data

Second rule of data viz – almost never use a pie chart…Hard to detect differences in categories visually, can easily become cluttered when categories exceed 4 or 5

Two variables – contingency tables count two variables using 311 dataset – add borough variable to columns

Marginal distribution – frequency distribution of each variable individually

Conditional distribution – distribution of a variable conditional on another. Ex, whats the distribution of eye condition for MALES. What % of Dept of Ed complaints come from Brooklyn.

When the distribution of one variable is the same for all categories of another variable, we say that they are independent. Ex, if percentage of females with each eye condition is the same as the percentage of females in the sample, we might say that gender is independent of eye condition. Ie, eye condition does not DEPEND on gender. There are robust tests for this that we’ll cover later in the semester, but just be aware that this is a big concept in analytics. Use caution when making claims of independence.

Segmented bar charts and mosaic plots? (remove Side-by-side is always better than stacked. Mosaic plots – probably not a great idea, but know they exist. Look at an example and try to make some general statement about it.

Simpsons paradox – use sales example. Who is the better salesperson? Think critically.

College admission example.

Homework questions: use another 311 dataset

1. What % of all reports came from Staten Island?
2. What % of all reports were for the Department of Transportation?
3. What % of reports from Manhattan were for the NYPD?
4. Come up with some question about the data and try to answer it with a pivot table and display it on a chart. Feel free to use filters to explore something you find interesting. Use your best judgement on the chart and explain your choices.