Stat 992 - Research Paper - Grammar of Graphics

Aaron Schram

Abstract

Introduction

The Grammar of Graphics (GoG) can be considered an orthogonal system of seven classes or data flow from which one can form a mathematical foundation for creating statistical graphics (Wilkinson, 2010). The core purpose of this data flow or class system is to design a grammar that can define any plot of any form. For further defining the GoG as a base language for forming statistical graphics, Wilkinson (2010) further describes this decomposition of the grammar on statistical graphics into seven strict categories which are the Variables, Algebra, Scales, Statistics, Geometry, Coordinates, and Aesthetics.

Then, we define Variables as the input of data from a file to a transformation of the data using table joins, filters, selects, and any other data pre-processing tools, which is denoted as the Algebra. From the Algebra, we have our the true data that we are interested in deriving a statistical graphic from, so we move onto the Scale class. Wilkinson (2010) denotes this the Scale class as the least observable or the most ignored class in with regards to the most common of charts. The core idea is that we a preferred scale or measurement for the data. These scales could consist of just an identity transformation on the algebra all the way up to non-linear transformations, but furthermore this class is also associated with the structure of the data. Here, the measurement or data structure just has to do with the class of the data itself, such as numeric, binary, interval, and so on. This truly marks the end of the data pre-processing phase as the remaining classes have to do with how the data will be mapped and plotted.

As for the last 4 classes, Statistics is the last class after all the data has been processed. It has to do with how we want to perceive the data. For example, do we want view every point as in a scatter plot, or do we want to view some kind of summary statistic as in histogram or bar chart. We even define this as output from regression lines or interactions plots from statistical tests and procedures. From the Statistics, the next class is the Geometry which is closely tied to the Statistics, since Geometry is all about how the data is being represented

on the plot itself. Geometry is the part where we specify if we are looking at lines, points, or bars, and how we perceive these on a graph. An easier way to examine this would be through the use of geom functions used in ggplot2 package (v3. 3.3; Wickham, 2016). These represent the visualizations of each layer and the representations of the Statistics (Wickham, 2009). In other words, the Geometry is the plot being created.

Finally, our last two classes are straightforward. Second to last, we have Coordinates, and this is just the coordinate grid system that we are plotting in. Then, we have the Aesthetic class which produces are final visual given all other classes. Wilkinson (2010) breaks the Aesthetic class into seven sub-functions based off of Bertin's *Visual Variables* (1967), and calls them position, size, shape, orientation, brightness, color, and granularity. Therefore, the Aesthetics class can be perceived as everything to do with the visualization of the graphic from the size each point to the color scheme of the graph itself.

Minimizing the Grammar

References