R Summer School - Graphics II

Jack Euesden

June 17, 2014

1 Visualising Data

For this part of the tutorial we will work with the diamonds dataset from the talk. This dataset maps a number of properties, including price, size and colour, for 53,000 diamonds - it is part of the ggplot2 package. Start by loading the ggplot2 package and reading the diamonds data. In order to make it easier to use, reduce the size of the dataset down to a random sample of 5,000 diamonds.

1.1 Histograms

- Plot histograms of depth, price, carat. What are the best binwidths for each of these variables?
- Optional: plot as densities and add smoothed lines to these objects
- Optional: split this into facets based on cut

1.2 Density Plots

- Produce a density plot of Price
- Split this density plot by diamond cut using different colours
- The densities overlap, and its hard to see what's going on. Edit the transparency to make the graph clearer. What is a sensible level of transparency to use? (hint – the factor 'cut' has 5 levels)

1.3 Scatter Plots

- Plot a scatter plot of price against carat
- Colour points by diamond colour
- Split into facets by cut

1.4 Pie and bullseye charts

- A pie chart is just a stacked bar chart on a polar axis. Plot a pie chart of diamond clarity by first using the bar geom, mapping clarity to fill to create a stacked bar graph. Then modify the co-ordinates of the plot to use a polar system.
- Now plot a bullseye chart by changing the axis along which clarity is displayed

2 Graphing Statistical Models

The economics dataset tracks US unemployment, personal consumption, personal savings etc over several decades in the 20th Century. Use the economics dataset to build up a plot in layers, experimenting with different ways of adding best-fit lines to a scatter plot:

- We are curious how unemployment changed over time: Create a geometric object for a line graph, plotting unemployment vs time
- Add a smoothed line line of best fit using the geom smooth.
- You can map other variables to this best fit line - try fitting a linear model for unemployment $\sim$ year, to the data, and adding a best fit line
- What confidence intervals have been used to plot the smoothed line? (hint: use ?stat_smooth), re-plot using 90% CI, and plot this curve in red.
- In the example above, why doesn't colour go within the aes() function?
- Create a stat object that can be added on to any ggplot to fit a y $\sim$ x linear best fit line without confidence intervals

3 EXTENSION: Using Maps

(liberally adapted from

`http://uchicagoconsulting.wordpress.com/tag/r-ggplot2-maps-visualization/`)

- We are interested in the sizes of colleges in the Midwest of the USA. First, plot a map of the USA using maptools. To do this, we need to read in a package called maps
- We can use the map_data function within ggplot to convert the data on US states, called "state", into a dataframe for analysis by ggplot.
- We can now use the geom_polygon to plot the US states based on the information in the states dataset. Map the information on US states to polygons - remember lat (latitude) and long (longitude) map to x and y and the group aesthetic maps states - these are also called group in the dataframe.
- Note that the geometric object polygon is used to superimpose a series of objects (48) based on their outlines. Now subset these to include only states in the Midwest (IA, IL, IN, KY, MI, MN, MO, OH and WI)
- Now plot the Midwest by mapping this reduced dataset to a polygon
- We have a separate dataset on college sizes across Midwest states – read this in, re-plot the Midwest states and add a layer to indicate the location of colleges - a point geom would work well
- Display size of enrolment using the size aesthetic
- Add names to each college - we use the text geom for this. The non-aesthetic parameters hjust and vjust might also be useful, to make sure the text isn't superimposed directly on top of the point indicating each college.
- OPTIONAL: there are many colleges in Chicago – can you offset the longitude and latitude of these points to make it clearer? Hint - use geom_jitter to add some random noise to each college's latitude and longitude.
- Finally, colour points by state