Introduction to ggplot2 Workshop

UVA Research Data Services: StatLab
Caitlin I. Steiner

March 18, 2015

ggplot2 package

"The emphasis in ggplot2 is reducing the amount of thinking time by making it easier to go from the plot in your brain to the plot on the page." (Wickham, 2012)

"Base graphics are good for drawing pictures; ggplot2 graphics are good for understanding the data." (Wickham, 2012)

Advantages of ggplot2

- consistent underlying grammar of graphics
- very flexible
- theme system for polishing plot appearance
- mature and complete graphics system

```
install.packages("ggplot2")
library("ggplot2")
```

Purpose of Layers I

1. To display the data (when starting off usually the only layer)
   - determine overall structure
   - look at local structure
   - notice outliers
2. To display a statistical summary of the data
   - display model predictions in the context of the data
   - showing the data helps us improve the model and showing the model helps reveal subtleties of the data we might otherwise overlook

Purpose of Layers II

3. To add additional metadata, context, and annonations
   - displays background context or annonations that help to give meaning to the raw data
   - could highlight important features of the data

Types of Layers I

Data The raw data that you want to plot. Must be in a **dataframe** object.

Geometric Object `geom_`

The visual elements or geometric shapes that will represent the data (e.g., points, lines, polygons, etc.)

Aethetics Mapping `aes()`

Aesthetics control the appearance of the geometric and statistical objects (e.g., color, size, shape, transparency, etc.)

Types of Layers II

Scales `scale_` Maps between the data and the aesthetic dimensions (e.g., color scales, size scales, shape scales, etc.)

Statistical Transformations Transforms the data, typically by summarizing it in some manner.

Faceting Describes how data is split into subsets and displayed as multiple small graphs

Coordinate System Describes the 2-D space that the data is projected onto (e.g. Cartesian coordinates, polar coordinates, map projections, etc.)

Types of Layers III

Position Adjustments Applies minor tweaks to the position of elements within a layer. Usually applied to discrete data and consists of 5 adjustments

1. `dodge` - positions objects so that there is no overlap at the side
2. `fill` - stacks objects on top of one another
3. `identity` - no position adjustment
4. `jitter` - jitter points to avoid overplotting
5. `stack` - stack overlapping objects on top of one another so that largest objects are at the back and smallest are at the front

Graph Object

- The basic idea behind the code structure is you specify different parts of the plot and add them together using the + operator.
- One first creates the object that specifies the plot by calling the function `ggplot()` (grammar of graphics plot).
- A graph object is not displayed until it contains at least one layer.

A general version of this command might look like

```
myGraph <- ggplot(<data>, aes(<x-variable>, <y-variable>))
```

Aesthetics I

Aesthetics refers to "something you can see", such as

- position
- color ("outside" color)
- fill ("inside" color)
- shape (of points)
- linetype
- size

These aesthetics are set by the `aes()` function, in which each geometric object only takes a subset of all possible aesthetic attributes.

Aesthetics II

- It is possible to declare certain aesthetics for the plot as a whole, instead of layer by layer.
  - Added layers inherit default aesthetics from the graph object.
  - Entire plot aesthetics can also be added, overwritten, or removed by re-specifying any of them for an individual layer.

Aesthetics III

There are two ways in which an aesthetic can be specified.

1. specific value (e.g. "Red")
   - don't specify within the `aes()` function
2. function of a variable (e.g. displaying data from different experimental groups in different colors)
   - specify within the `aes()` function so that it may vary

Geom Objects I

- A plot must consist of at least one geometric object.
  - There is no limit to how many geometric objects a plot can consist of.
- Each geometric object is added to a plot by using the + operator.
- Shortcut syntax format `geom_<object type>()`
- Some objects require certain aesthetics to be declared, while others are optional.
  - Bare minimum require is to declare the variable(s) that the geom represents

Geom Objects II

A list of all available geometric objects can be found by running

1. `help.search("geom_", package = "ggplot2")`
2. or typing `geom_` into the console of RStudio and then hitting Tab.

Stats & geoms I

Some geometric objects require you to generate the data necessary to plot them.

- can either enter in the values needed for the geom
- or use the ggplot2 built-in functions called 'stats'
  - used by a geom
  - or directly create a visual element on a layer of a plot

It is worth knowing about them because they enable you to adjust the properties of a plot.

Stats & geoms II

Common built-in stats are

- bin
- boxplot
- density
- qq
- smooth
- summary

Facets

If we want to compare groups of data alongside one another, we can

- map a discrete variable to an aesthetic (e.g., x position, color, or shape)
- create a subplot for each group and draw the subplots side by side

Facets

If we want to compare groups of data alongside one another, we can

- map a discrete variable to an aesthetic (e.g., x position, color, or shape)
- create a subplot for each group and draw the subplots side by side

We can use

`facet_grid` where we specify a variable to split the data into vertical subpanels, and another variable to split into horizontal subpanels

`facet_wrap` where the subplots are laid out horizontally and wrap around (default is square fitting all)

Overplotting

Sometimes with large data sets, you may have points that obscure one another.

To prevent this overplotting one can

- Make the points semitransparent by using the `alpha` aesthetic
- Bin the data into rectangles and map the density of the points to the fill color
- Bin the data into hexagons
- Use boxplots

References I

http://docs.ggplot2.org/current/

http://opr.princeton.edu/workshops/201501/

http://www.amazon.com/gp/product/0387981403/ref=as_li_ss_tl?ie=UTF8&camp=1789&creative=390957&creativeASIN=0387981403&linkCode=as2&tag=learningasyougo-20

http://www.r-bloggers.com/a-simple-introduction-to-the-graphing-philosophy-of-ggplot2/

References II

http://projects.iq.harvard.edu/rtc/event/r-graphics

file:///C:/Users/stei%20stei/AppData/Local/Temp/Temp1_Rgraphics.zip/Rgraphics/Rgraphics.html#outline-

``Discovering Statistics Using R" by Andy Field, Jeremy Miles, and Zoe Field

``R Graphics Cookbook" by Winston Chang