A Two Page Guide to ggplot2

Andy Grogan-Kaylor (http://www.umich.edu/~agrogan)
June 26, 2019

[Share on Twitter]

1 Why?

A great deal of data analysis and visualization involves the same core set of steps: get some data, clean it up a little, run some descriptive statistics, run some bivariate statistics, create a graph or a visualization. ggplot can be an important part of a replicable, automated, documented workflow for complex projects.

have a question → get data → process and clean data →
visualize data → analyze data → make conclusions

Given the fact that we often want to apply the same core set of tasks to new questions and new data, there are ways to overcome the steep learning curve and learn a replicable set of commands that can be applied to problem after problem.

2 The Essential Idea Of ggplot2 Is Simple

There are 3 essential elements to any ggplot call:

1. An aesthetic that tells ggplot which variables are being mapped to the x axis, y axis, (and often other attributes of the graph, such as the color fill). Intuitively, the aesthetic can be thought of as what you are graphing.
2. A geom or geometry that tells ggplot about the basic structure of the graph. Intuitively, the geom can be thought of as how you are graphing it.
3. Other options, such as a graph title, axis labels and overall theme for the graph.

3 Get Started

library(ggplot2)  # beautiful graphs

library(ggthemes)  # nice themes for ggplot2

<table>
<thead>
<tr>
<th>predictor</th>
<th>outcome</th>
<th>group</th>
</tr>
</thead>
<tbody>
<tr>
<td>96.59</td>
<td>97.31</td>
<td>A</td>
</tr>
<tr>
<td>90.43</td>
<td>93.97</td>
<td>A</td>
</tr>
<tr>
<td>101.9</td>
<td>93.98</td>
<td>A</td>
</tr>
<tr>
<td>103.8</td>
<td>108.2</td>
<td>A</td>
</tr>
<tr>
<td>114.9</td>
<td>112.4</td>
<td>A</td>
</tr>
<tr>
<td>106</td>
<td>104.5</td>
<td>A</td>
</tr>
</tbody>
</table>

Figure 1: Sample Data
4 Some Examples

4.1 One Continuous Variable

```r
# anything that starts with a '#' is a comment

ggplot(mydata, # the data I am using
       aes(x = outcome)) + # the variable I am using
       geom_histogram() # how I am graphing it
```

4.2 One Categorical Variable

```r
ggplot(mydata, # the data I am using
        aes(x = outcome,
            fill = group)) + # the variable I am using
        geom_bar() # how I am graphing it
```

4.3 Continuous by Continuous

```r
ggplot(mydata, # the data I am using
        aes(x = predictor, y = outcome,
            color = group)) + # the variables I am using
        geom_point() + # how I am graphing it
        geom_smooth() +
        labs(title = "My title",
             x = "title for x axis",
             y = "title for y axis") +
        scale_colour_manual(name = "Group",
                             values = c("#FF0099",
                                         "#0099FF"), # manual colors
                             labels = c("group1",
                                         "group2").
```

4.4 Add Some Options

```r
ggplot(mydata, # the data I am using
        aes(x = predictor, y = outcome,
            color = group)) + # the variables I am using
        geom_point() + # how I am graphing it
        geom_smooth() +
        theme_minimal() + # theme
        theme(plot.title = element_text(size = 20, # add to theme
                                         color = "#0099FF"))
```

5 Changing variables from factor to numeric (e.g. `aes(x = as.numeric(outcome)))`, and vice versa can sometimes be a simple solution that solves a lot of problems when you are trying to graph your variables.