The Versatility of GGPlot

Ellie Sun

10/26/2017

About GGPlot

GGPlot2 is a package in R that helps streamline the process of creating plots, graphs, charts, and other visuals by utilizing data in the form of a data frame. It loads all the functions associated with the graphing function ggplot() and all the geom() options to create a variety of graphics. Being able to generate cohesive and useful visuals is an important skill becaue it enhances presentations and helps the viewer digest information becuase it is presented in a differnt medium. It is a tool that helps organize and analyze data so people do not have to spend their time staring at lists of numbers that are hard to comprehend and take time to digest.

The Origin of GGPlot

GGplot is based on the grammar of graphics to create plots, hence the name - ggplot. Now what is “the grammer of graphics” you ask? It was an idea developed by Leland Wilkinson, a statistician and computer scientist, that suggests that a grammar should be implemented to make all aspects of creating and using a graphic simple. Hadley Wickham of Iowa State University followed that idea to create ggplot and won the John Chambers Statistical Software Award from The Statistical Computing Section of the American Statistical Association for this package in 2006. Wickham used a format that starts simple and adds on the extra components of graphs in a consistant manner. It is an easy way to implement various graphical aspects to a data set with code that is easy to understand and less complex than graphing all aspects separately.

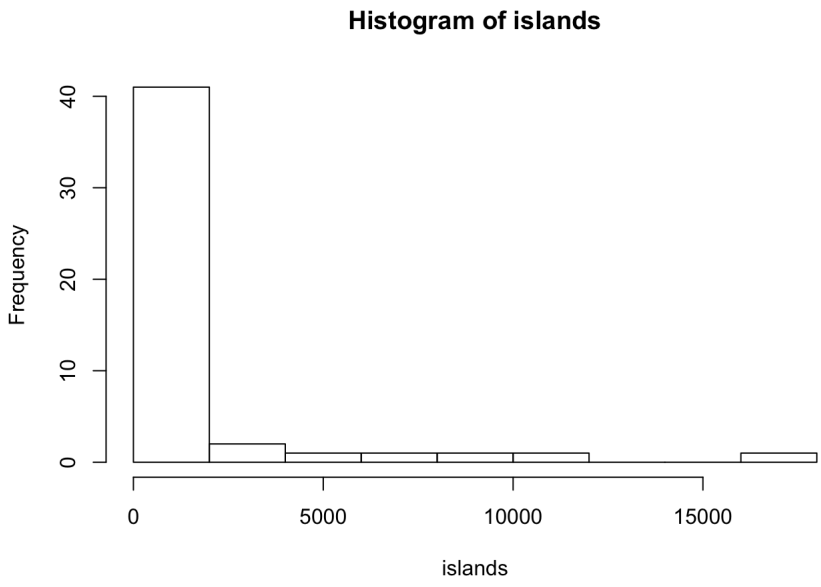
GGPlot is part of Tidyverse, a collection of packages for R that are designed for data science. These packages are consistantly being maintained and updated to provide the best resources for data sceince for free.

The Advantages of GGPlot

First, lets look at some code involved with generating plots.

Here we have a basic plot that is made with a base code function (like barplot() or hist()).

```
hist(islands)
```

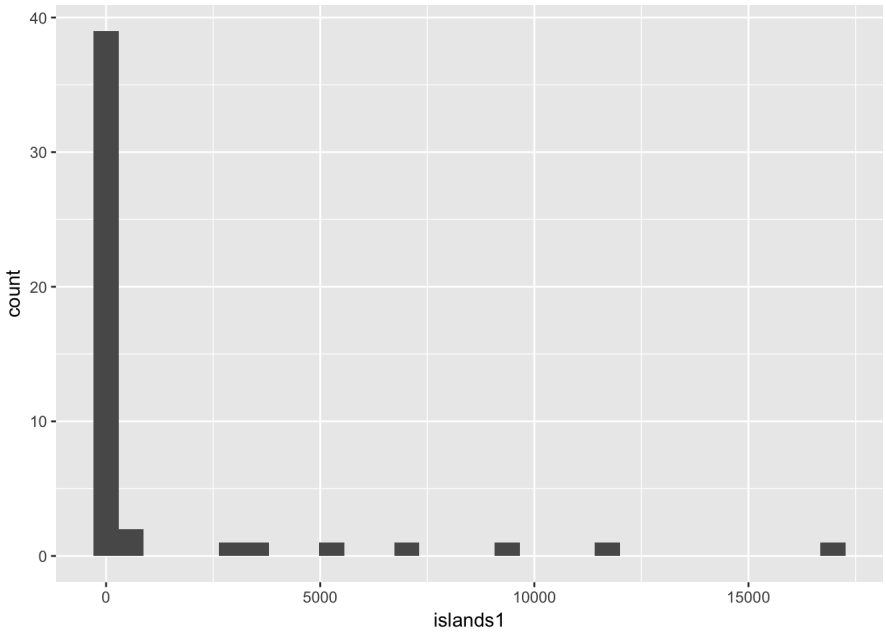


Now lets recreate that plot using ggplot.

```
#Please note that the produced plot is basic and missing titles to show that the hist() function is more user frie
ndly for simple and straightforward histograms that depicts basic data without many components.
library(ggplot2)
islands1 <- data.frame(islands)
ggplot(islands1, aes(x = islands1)) +
  geom_histogram()
```

```
## Don't know how to automatically pick scale for object of type data.frame. Defaulting to continuous.
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

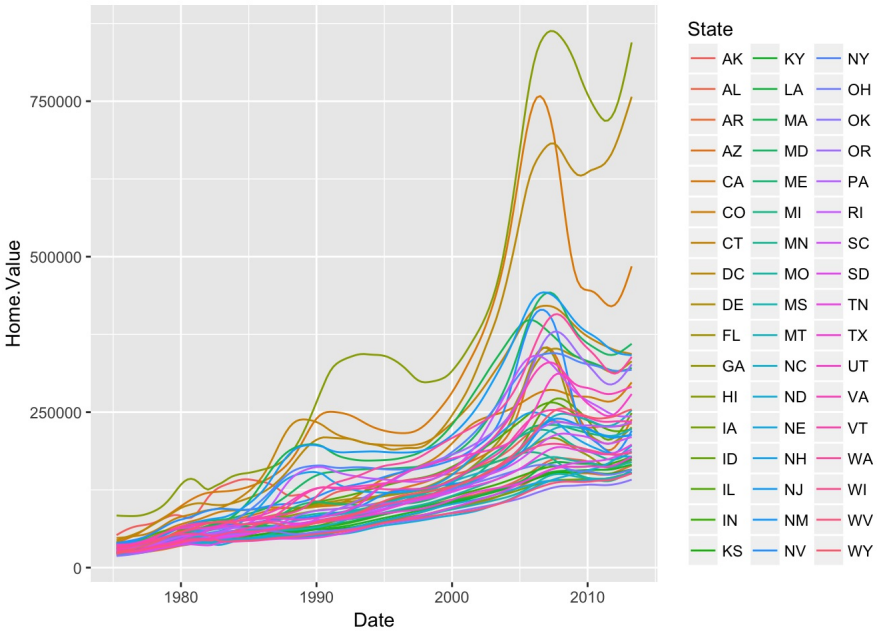


With these two plots, we can see that using the base graphics of R is easier when you need a simple graph or when using data that is not in a data frame. The base function autogenerates a title while ggplot would require adding the title manually. In addition to the premade title, the format of the base histogram is more condensed and easier to read.

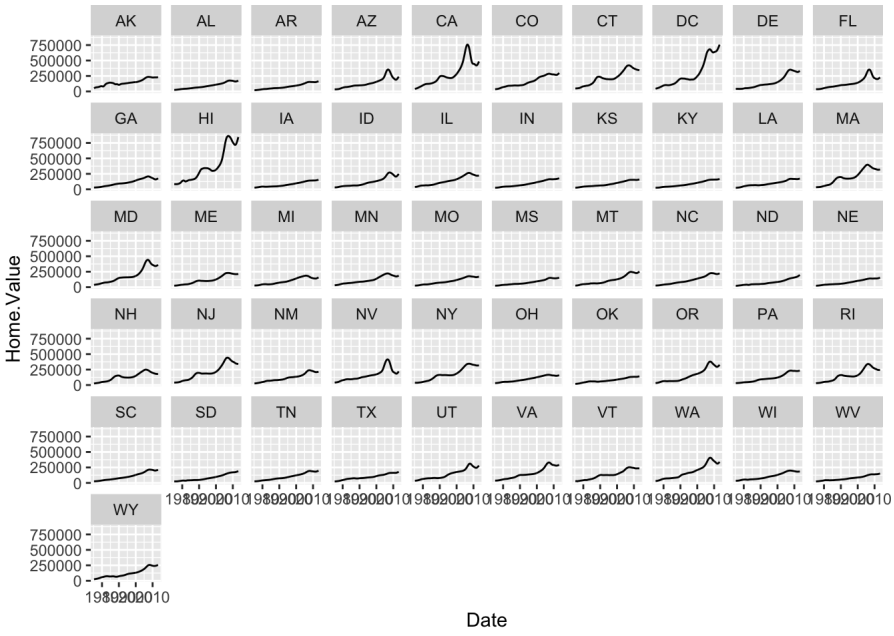
However, with more complex graphics that look at multiple aspects of data, ggplot is a much better resource. It allows you to generate a theme to manipulate appearance, like labels of axises and facets as well the plot background and the legend. It can also generate scales to show change over time or the significane of certain findings by using a color scale that indicates those aspects, or whatever aspects desired. This allows you to present more information about the data you are analyzing in a more streamlined manner because through ggplot, you are able to visually present many components of the data.

Below are some examples of utilizing theme and scale through ggplot. (Courtesy of Harvard)

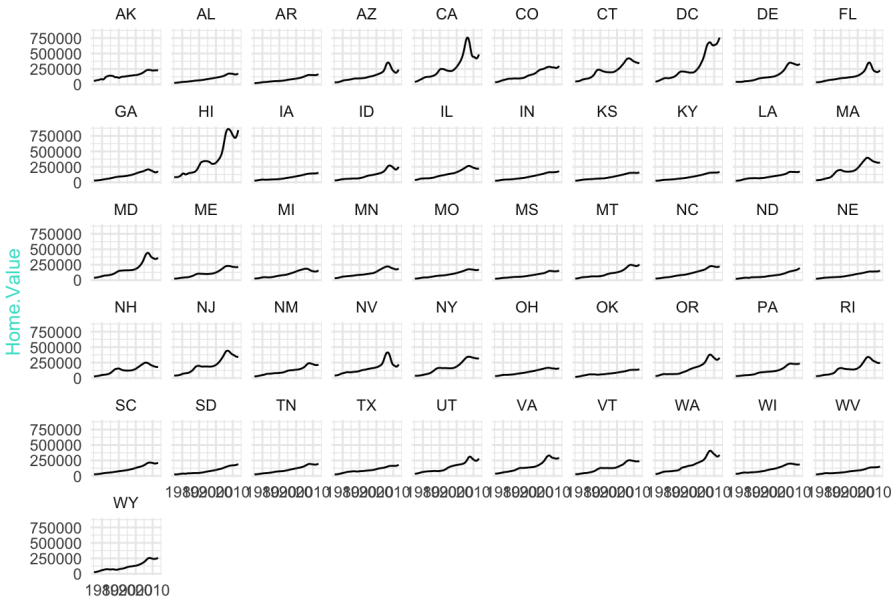
```
#Loading data about housing in the United States.
housing <- read.csv("dataSets/landdata-states.csv")

#Showing housing trends in the United States.
#See how ggplot can take the data for each state and present it in a visually appealing and clear way.
p5 <- ggplot(housing, aes(x = Date, y = Home.Value))
p5 + geom_line(aes(color = State))
```



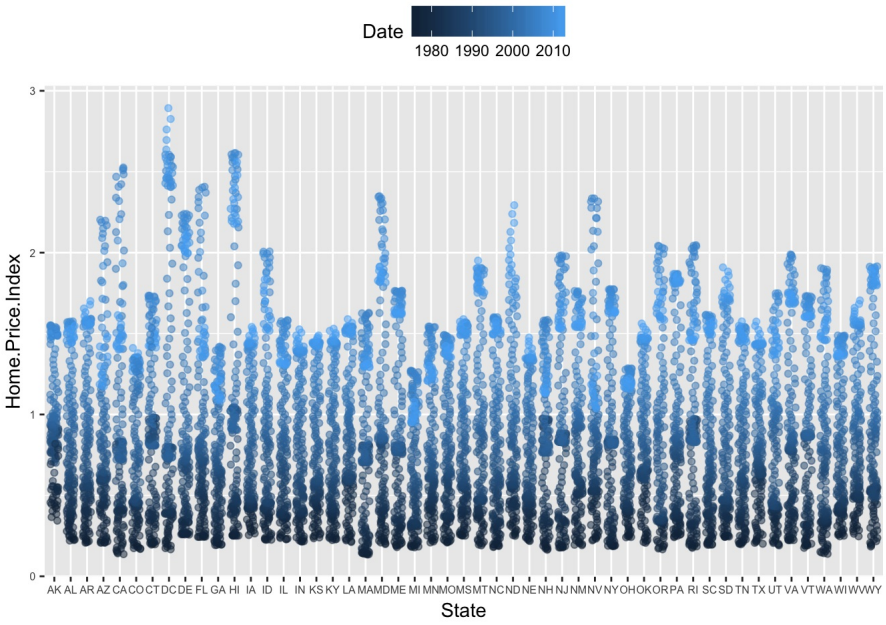

```
#Using Theme
(p5 <- p5 + geom_line() +
   facet_wrap(~State, ncol = 10))
```




```
p5 + theme_minimal() +
  theme(text = element_text(color = "turquoise"))
```

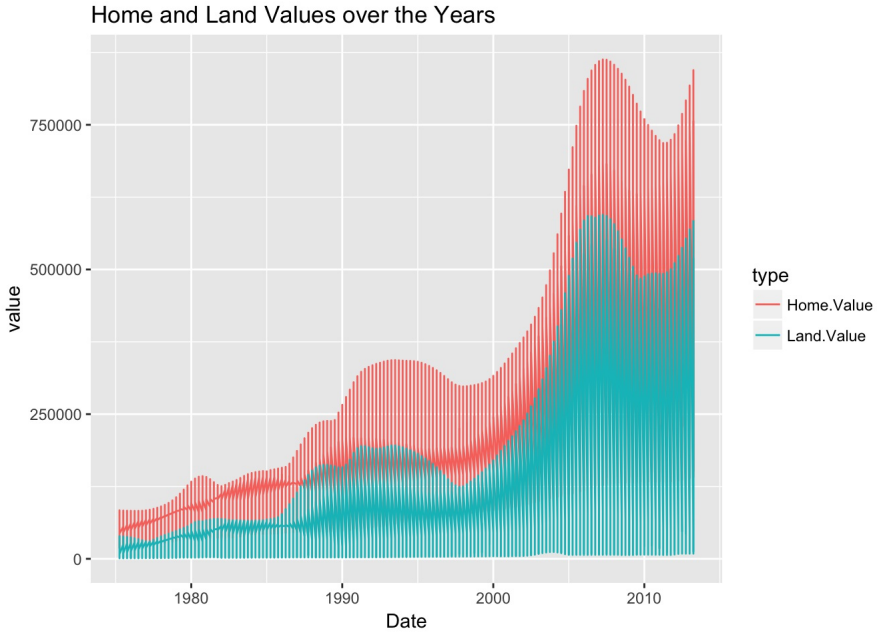



```
#Examples of changing scale
p3 <- ggplot(housing,
             aes(x = State,
                 y = Home.Price.Index)) +
       theme(legend.position="top",
             axis.text=element_text(size = 6))
(p4 <- p3 + geom_point(aes(color = Date),
                       alpha = 0.5,
                       size = 1.5,
                       position = position_jitter(width = 0.25, height = 0)))
```



GGplot also allows you to easily create layers of data.

```
library(tidyr)
home.land.byyear <- gather(housing,
                           value = "value",
                           key = "type",
                           Home.Value, Land.Value)
ggplot(home.land.byyear,
       aes(x=Date,
           y=value,
           color=type)) +
  geom_line() +
  labs(title = "Home and Land Values over the Years")
```



These are among the few benefits of using ggplot. It also allows you to change the shape or labels of data points. use continous values for y to generate more complex graphs, visualize error or range in vairous orientations, display your data with three axises, utilize maps and display information using map coordinates.

Conclusion

As seen above, GGplot is a useful tool in data science because it is an efficient way to generate visuals of large data sets to make it easier to comprehend. Being able to analyze, organize and present data is a vital skill because so many jobs are now focused on looking at data and being able to make an educated decision based on the data. Being able to see the various aspects of data and use different perspectives helps develop a deeper understanding of the data. It is versatile, simple, free and always improving!

References

http://tutorials.iq.harvard.edu/R/Rgraphics/Rgraphics.html#the_1_faq

http://had.co.nz/ggplot2/

http://ggplot2.org/resources/2007-vanderbilt.pdf

http://ggplot2.org/resources/2007-past-present-future.pdf

http://stat-computing.org/awards/jmc/winners.html

http://ggplot2.tidyverse.org/

https://www.tidyverse.org/packages/

http://seananderson.ca/ggplot2-FISH554/

https://www.youtube.com/watch?v=rsG-GgR0aEY

https://github.com/ucb-stat133/stat133-fall-2017/blob/master/cheat-sheets/ggplot2-cheatsheet-2.1.pdf