

Data Visualisation with R using ggplot

Chipo Zidana

Department of Mathematics and Statistical Sciences, **BIUST**

What is Data visualisation?

- **Data visualisation** practice of translating information into a visual context
- a step in data science process
- It uses visual elements like **charts, graphs, dashboards, info graphics and maps**
- **Goal:** identify patterns, trends and outlier in data
- **Big data is here!!!** call for better tools

Tell your story right

- Marry the data and the visuals
- delicate balancing act between form and function
- its an art of great analysis and story telling

Who data Visualisation?



ME,ALL

AND YOU

Data Talks



Why Data Visualisation

- “A good sketch is better than a long speech.”
- Sketching out our data by visualizing is more impactful than simply describing the patterns and trends we find.
- “The simple graph has brought more information to the data analyst’s mind than any other device.” — **John Tukey**
- it’s easier to learn from something that we can see rather than read
- **We need a powerful tool!!** In R data visualisation is a snap using
 - 1 Base R
 - 2 Lattice
 - 3 Grid
 - 4 ggplot2

Why ggplot2

- Uses the grammar of graphics - Easy and layers tell the complete story
- “It can do quick-and-dirty and complex”-Mendy Mejia
- Easy superposition, facetting, Automatic legends, colors, etc.ie Multivariate exploration is greatly simplified.
- Store any ggplot2 object for modification or future recall. Super useful for packages.
- Lots of users (less bugs, more help on Stack Overflow).
- Lots of extensions.ggtheme, ggrepel etc
- flexible- Nice saving option
- Plots can evolve and devolve with minimal changes

HOW in R?

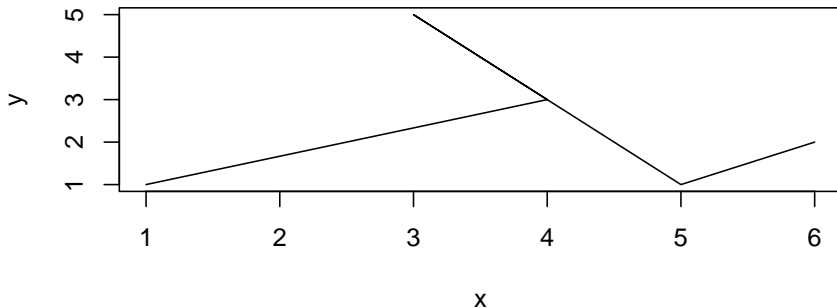
```
#install.packages("tidyverse")  
library(ggplot2)
```

What does ggplot stand for?

- A Grammar of graphics
- A grammar of graphics defines the rules of structuring mathematics and aesthetic elements into a meaningful graph.
- Leland Wilkinson (2005) designed the grammar upon which ggplot2 is based.
- **ggplot2** package follows and describes data in the graph aspects

General aspects/What is in Graph?

- there are many type of visualisations but we can generalise as follows



GGplot2 Grammar aspects/Layers

- ❶ **Data:** variables mapped to aesthetic features of the graph.
- ❷ **Geoms:** objects/shapes on the graph.
- ❸ **Stats:** statistical transformations that summarize data,(e.g mean, confidence intervals).
- ❹ **Scales:** mappings of aesthetic values to data values. Legends and axes visualize scales.
- ❺ **Coordinate systems:** the plane on which data are mapped on the graphic.
- ❻ **Faceting:** splitting the data into subsets to create multiple variations of the same graph (paneling).
- **ANY Plot can be described uniquely as a combination of these aspects parameters with layers.**

ggplot2 Aspects: Plot = data + Aesthetics + Geometry

- *Data* -data frame/table
- *Aesthetics* - indicate x and y variables, control colour, size and shape
- *geometry* - defines the type of graphics (histogram, box plot, line plot, density plot, dot plot,)

```
-ggplot(data =< DATA >) +  
  < GEOM FUNCTION >(  
    mapping = aes(< MAPPINGS >),  
    stat =< STAT >,  
    position =< POSITION >) +  
  < COORDINATE FUNCTION > +  
  < FACET FUNCTION >
```

DATA

- Data should be tidy/ clean (**Use dplyr to re arrange data tidy -disadvantage**)
- Tidy data is easier to work and work with ie:
 - ① Each variables are represent a column
 - ② each row present an observation
 - ③ each value has a unique column and row
- import and tidy your data before visualising;
- **DATA CLEANING IS VITAL**

TIDY DATA

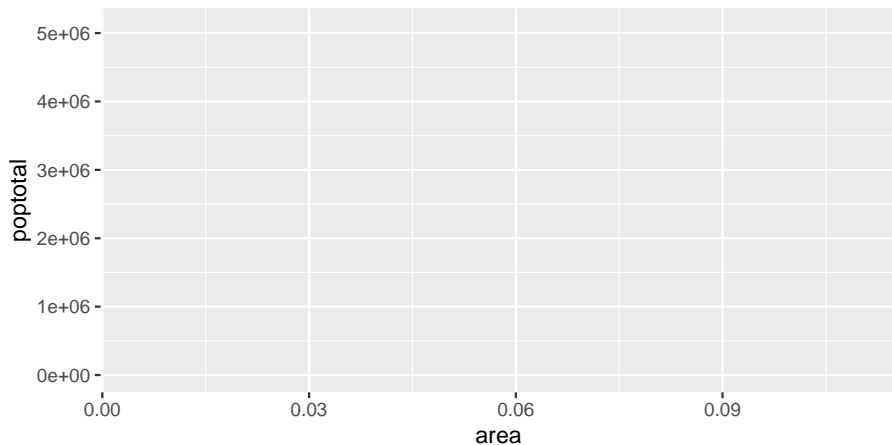
```
## # A tibble: 437 x 28
##       PID county  state  area poptotal popdensity popwhite p
##   <int> <chr>   <chr> <dbl>    <int>      <dbl>    <int>
## 1   561 ADAMS    IL    0.052   66090      1271.    63917
## 2   562 ALEXAN~ IL    0.014   10626       759     7054
## 3   563 BOND     IL    0.022   14991       681.    14477
## 4   564 BOONE    IL    0.017   30806      1812.    29344
## 5   565 BROWN    IL    0.018    5836       324.     5264
## 6   566 BUREAU   IL    0.05    35688       714.    35157
## 7   567 CALHOUN IL    0.017    5322       313.     5298
## 8   568 CARROLL IL    0.027   16805       622.    16519
## 9   569 CASS     IL    0.024   13437       560.    13384
## 10  570 CHAMPA~ IL    0.058  173025      2983.   146506
## # ... with 427 more rows, and 19 more variables: popasian <int>,
## #   popother <int>, percwhite <dbl>, percblack <dbl>, percasian <dbl>,
## #   percother <dbl>, popadults <int>, perc...
```

The ggplot() structure

- In the ggplot() function - specify the data set that holds the variables we will be mapping to aesthetics, the visual properties of the graph.
- The data set must be a data.frame object.
- **ggplot(data, aes(x=xvar, y=yvar))**
 - ① data- data frame
 - ② x and y: aesthetics that position objects on the graph
 - ③ xvar and yvar: names of variables in data mapped to x and y
- **NB:** aes are specified inside aes(), which is itself nested inside of ggplot().
- aesthetics specified inside of ggplot() are inherited by subsequent layer

Global GGplot

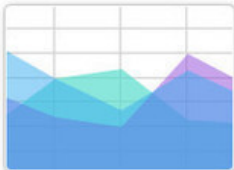
```
ggplot(data = midwest, aes(x = area, y = poptotal))
```



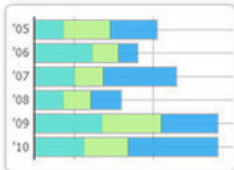
Add Geom function

- define geometries, - compute summary statistics, -
- define what scales to use, - or even change styles.

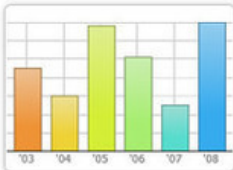
Area



Horizontal Bar



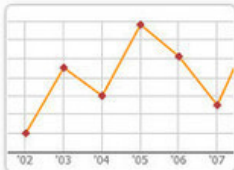
Vertical Bar



Donut



Line



Pie



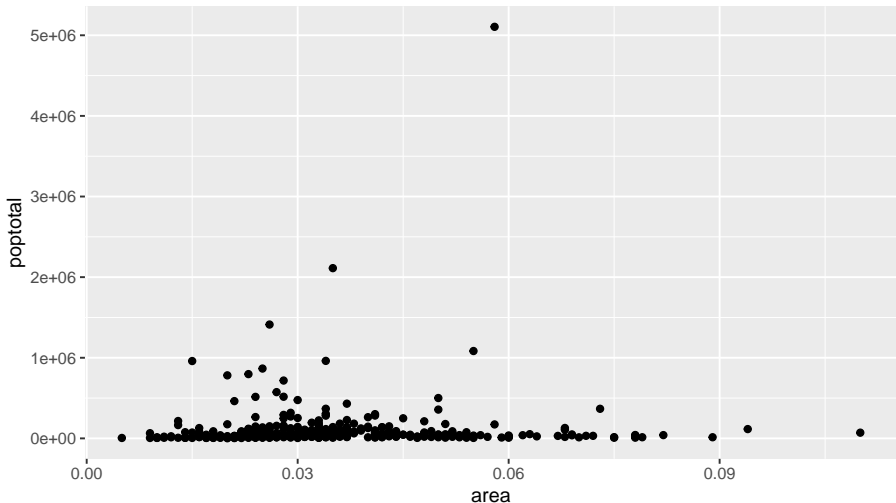
Radar

Scatter

Spline

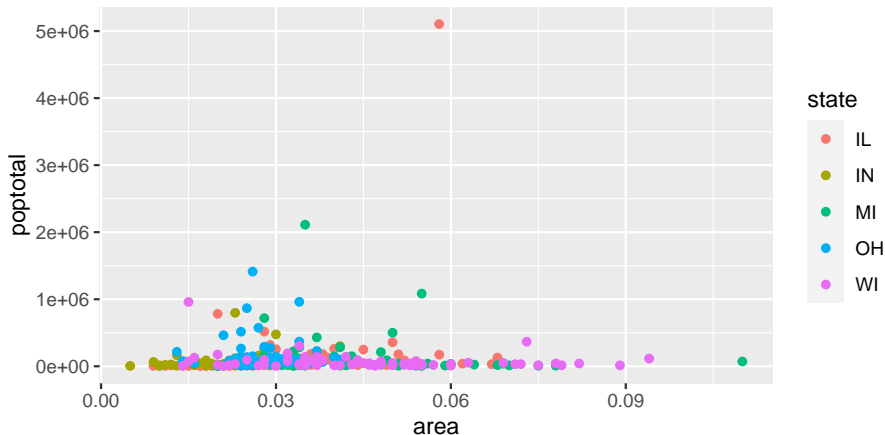
① Scatter Plot: `geom_point()`

```
ggplot(data = midwest, aes(x = area, y = poptotal))+  
  geom_point()
```



Add colour

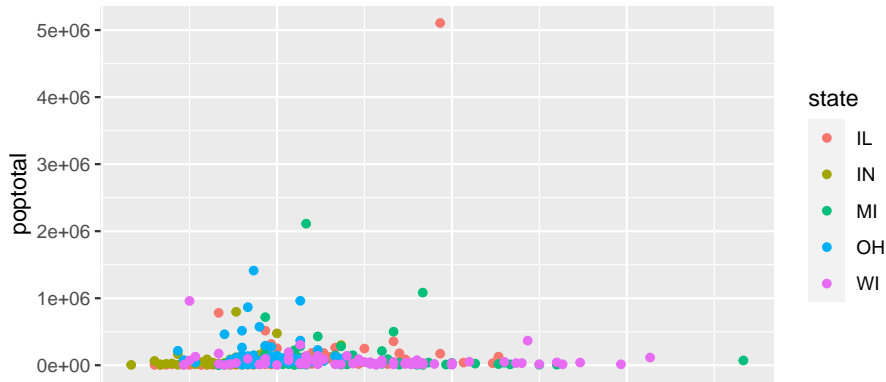
```
ggplot(data = midwest, aes(x = area, y = poptotal,  
  color= state)) + geom_point()
```



Custom the graph

- Name the ggplot plot and custom it

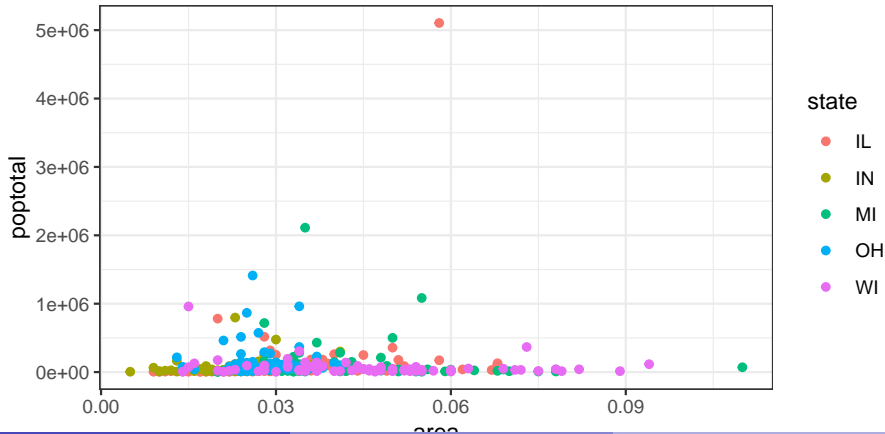
```
g = ggplot(data = midwest, aes(x = area, y = poptotal, color =  
  geom_point()  
plot(g)
```



Change background

- background themes can be changed to different ways:
- **eg:** `theme_bw()` / `theme_classic()`

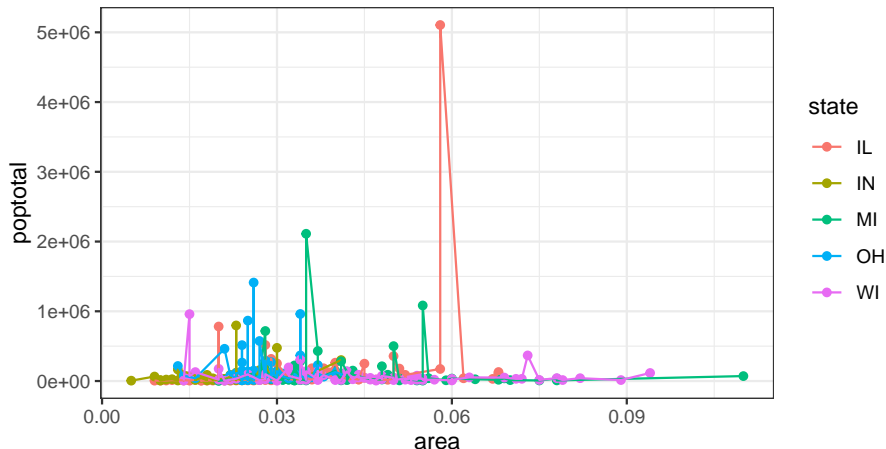
```
g+theme_bw()
```



Add Another Geom function

② Line graph: `geom_line()`

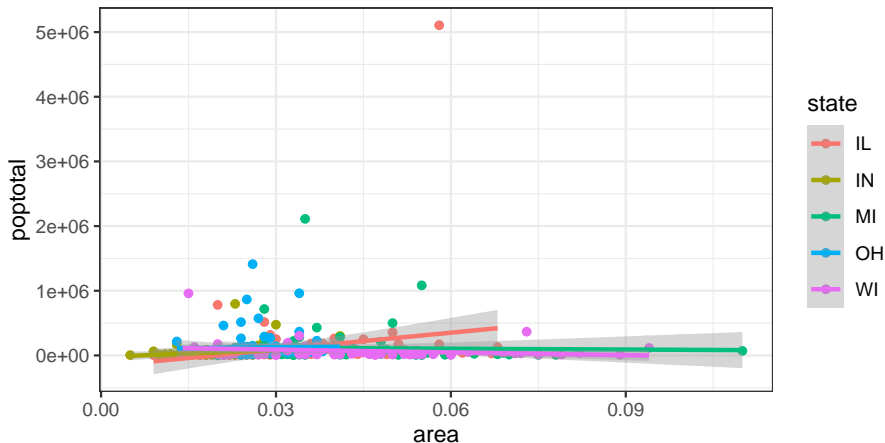
```
g+theme_bw()+geom_line()
```



Add Geom Smooth Layer

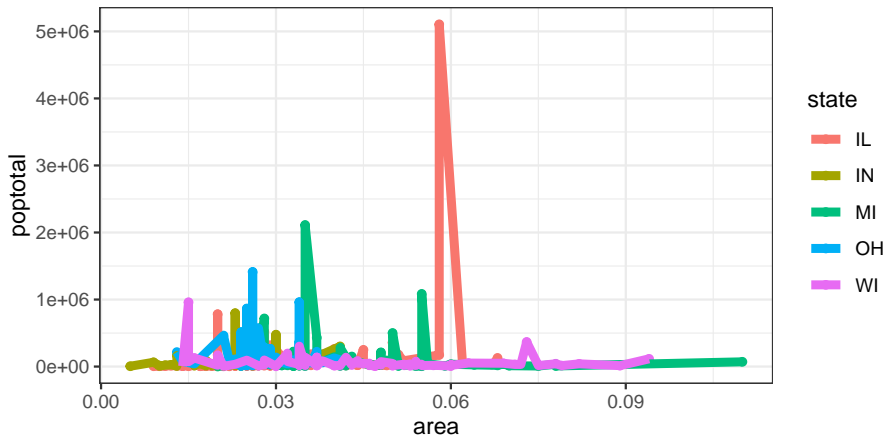
```
g+theme_bw()+geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



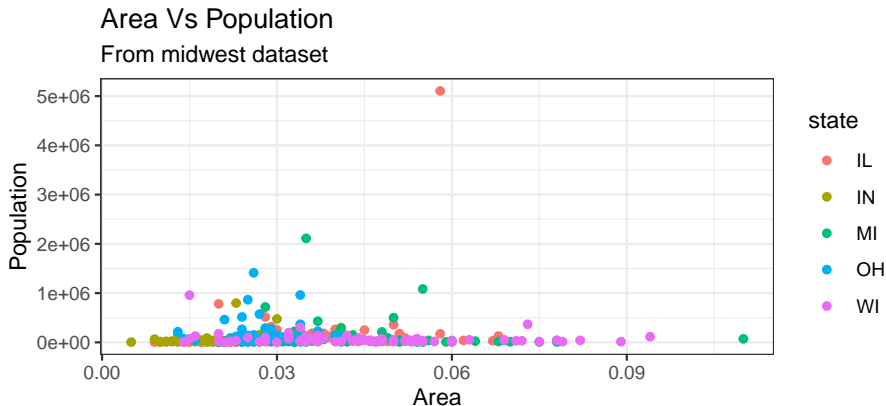
Change size of points and lines

```
g+theme_bw()+geom_line(size =2)
```



Add Title and Axis labels

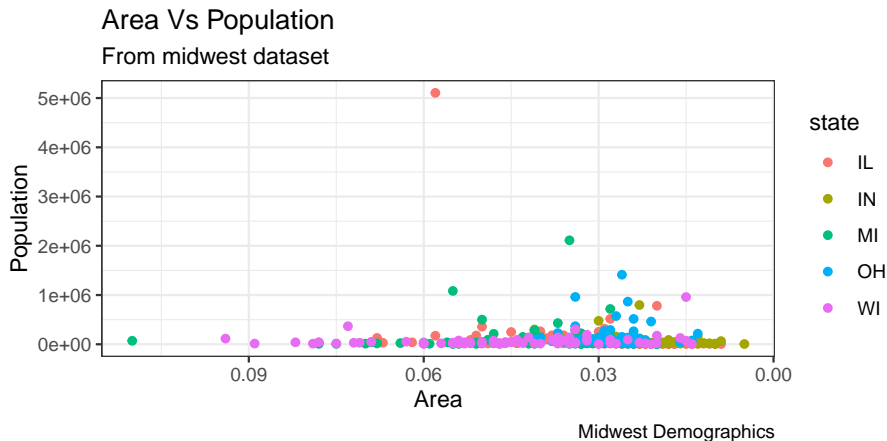
```
gg=g+theme_bw()+ labs(title="Area Vs Population",  
  subtitle="From midwest dataset",y="Population",  
  x="Area", caption="Midwest Demographics")  
gg
```



Scale reservesing

- If you need to reverse the scale, use `scale_x_reverse()`.

```
gg+ scale_x_reverse()
```

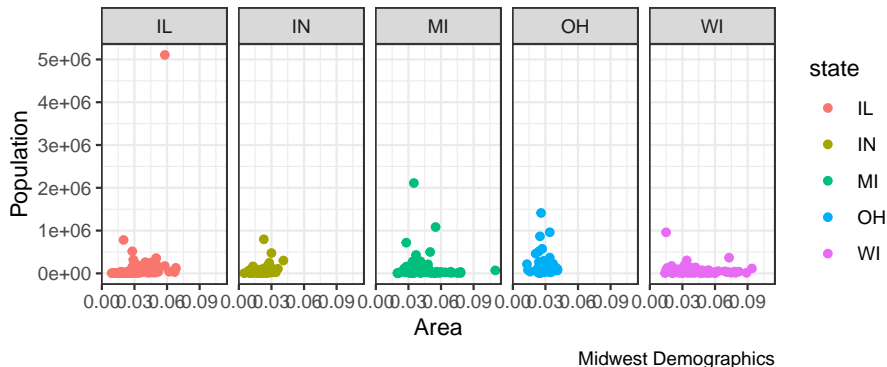


Facets

```
gg+facet_grid(~state)
```

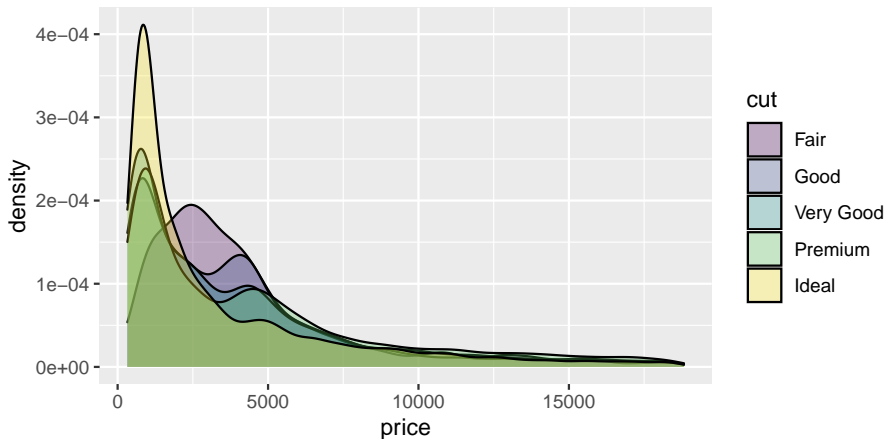
Area Vs Population

From midwest dataset



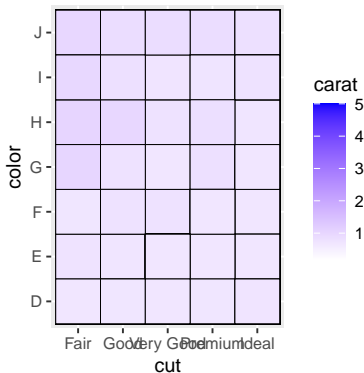
Flexibility

```
data(diamonds)
ggplot(data=diamonds, aes(x=price, fill=cut)) +
  geom_density(alpha=0.3)
```



Heat Map

```
ggplot(diamonds,aes(x = cut, y = color, fill = carat))+  
  geom_tile(color = "black") +  
  scale_fill_gradient(low = "white", high = "blue") +  
  coord_fixed()
```

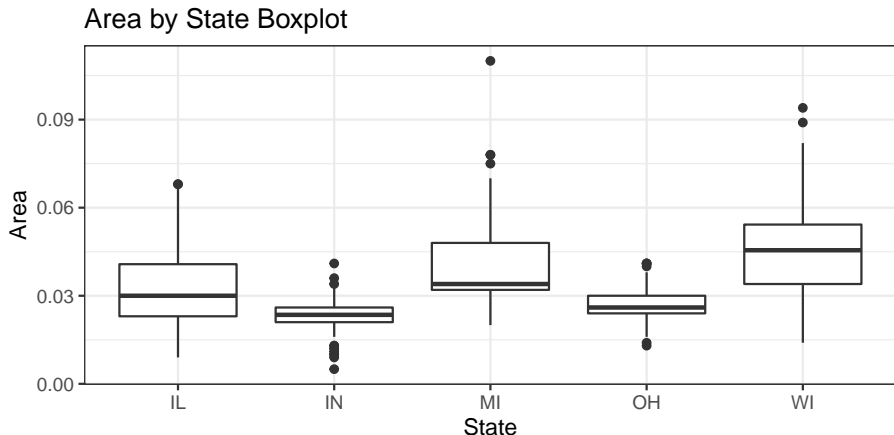


Section 1

Box Plot

geom_boxplot()

```
bx1 = ggplot(data = midwest, aes(x = state, y = area)) +  
  geom_boxplot()+theme_bw()+ labs(title = "Area by State Boxplot")  
bx1
```



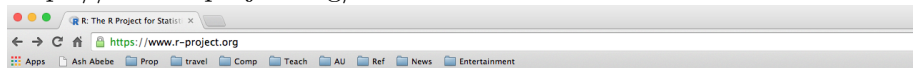
Resources

- ① Free R
- ② Free R studio
- ③ Many websites to

Learn and Learn and Learn More

Install R

<https://www.r-project.org/>



[Home]

Download

CRAN

R Project

[About R](#)
[Contributors](#)
[What's New?](#)
[Mailing Lists](#)
[Bug Tracking](#)
[Conferences](#)
[Search](#)

R Foundation

[Foundation](#)
[Board](#)
[Members](#)
[Donors](#)
[Donate](#)

Documentation

[Manuals](#)
[FAQs](#)
[The R Journal](#)
[Books](#)
[Qualifications](#)

The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To [download R](#), please choose your preferred [CRAN mirror](#).

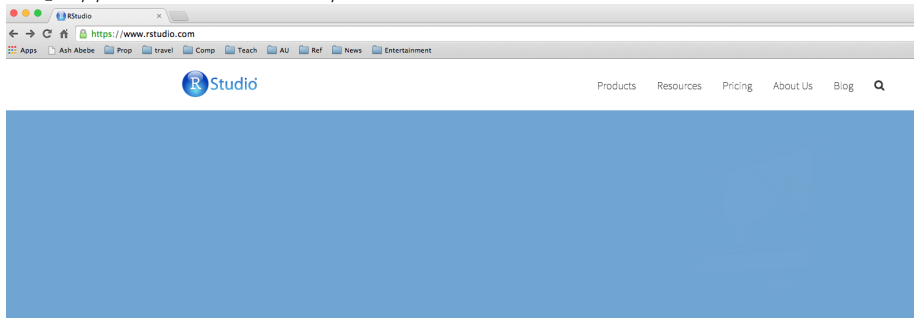
If you have questions about R like how to download and install the software, or what the license terms are, please read our [answers to frequently asked questions](#) before you send an email.

News

- [R version 3.2.2 \(Fire Safety\)](#) has been released on 2015-08-14.
- [The R Journal Volume 7/1](#) is available.
- [R version 3.1.3 \(Smooth Sidewalk\)](#) has been released on 2015-03-09.
- [useR! 2015](#), will take place at the University of Aalborg, Denmark, June 30 - July 3, 2015.
- [useR! 2014](#), took place at the University of California, Los Angeles, USA June 30 - July 3, 2014.

Install R-Studio

<https://www.rstudio.com/>



Powerful IDE for R

RStudio IDE is a powerful and productive user interface for R. It's free and open source, and works great on Windows, Mac, and Linux.

[Learn More >](#)



R Packages

Our developers and expert trainers are the authors of several popular R packages, including `ggplot2`, `plyr`, `lubridate`, and others.

[Learn More >](#)

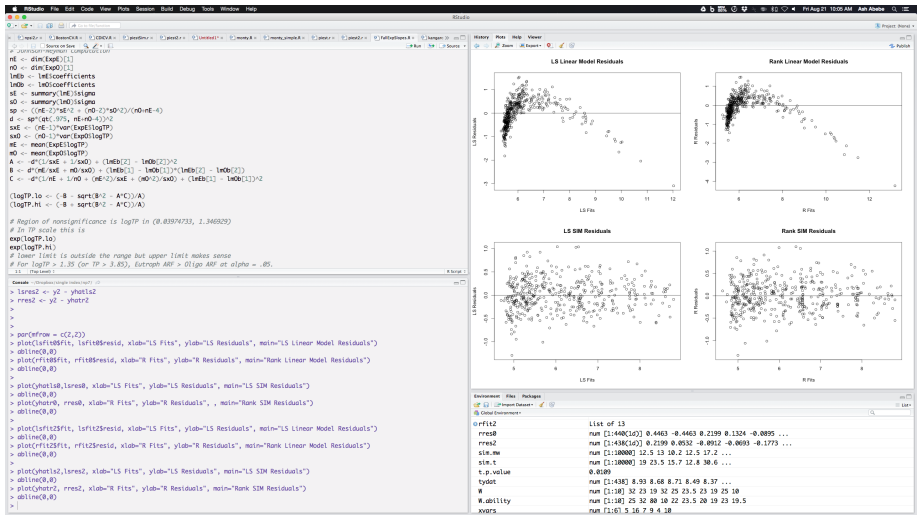


Bring R to the web

Shiny is an elegant and powerful web framework for building interactive reports and visualizations using R — with or without web development skills.

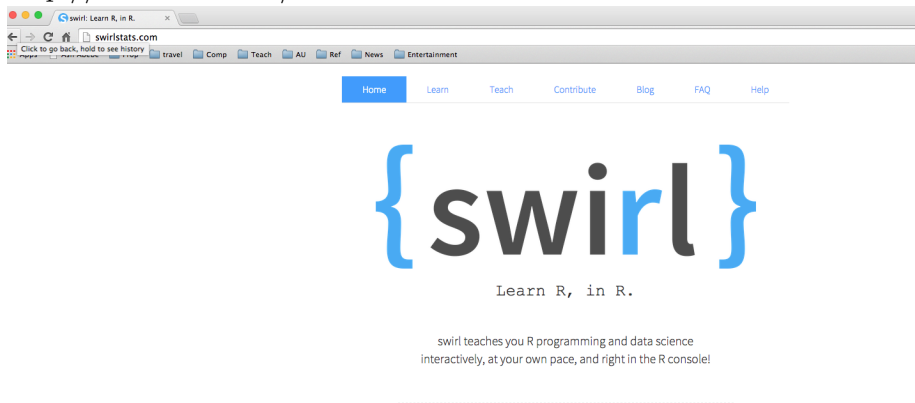
[Learn More >](#)

My R-Studio



Learn R

<http://swirlstats.com/>



<http://tryr.codeschool.com/>



CHAPTER 1

Try R

In this first chapter, we'll cover basic R expressions. We'll start simple, with numbers, strings, and true/false values. Then we'll show you how to store those values in variables, and how to pass them to functions. We'll show you how to get help on functions when you're stuck. Finally we'll load an R script in from a file.

Let's get started!

Try R is Sponsored By:

O'REILLY®



Complete to
Unlock

Expressions

1.1

Type anything at the prompt, and R will evaluate it and print the answer.

Let's try some simple math. Type the below command.

Learn R

https://www.datacamp.com/

Introduction to R | DataCamp

https://www.datacamp.com/courses/free-introduction-to-r

Apps Ash Abebe Prop travel Comp Teach AU Ref News Entertainment

DataCamp
We're hiring!

Courses Pricing Teams ash@auburn.edu

Course

Introduction to R

With over 2 million users worldwide R is rapidly becoming the leading programming language in statistics and data science. Every year, the number of R users grows by 40%, and an increasing number of organizations are using it in their day-to-day activities. In this introduction to R, you will master the basics of this beautiful open source language such as factors, lists and data frames. With the knowledge gained in this course, you will be ready to undertake your first very own data analysis.


Start Course

Or, take a subscription

- Difficulty:** Beginner
- Duration:** 4 hours
- Participating:** 81842

Chapter 1: Intro to basics

In this chapter, you will take your first steps with R. You will learn how to use the console as a calculator and how to assign variables. You will also get to know the basic data types in R. Let us get started!

 Number of exercises: 8



Course given by:

Jonathan Cornelissen
DataCamp



Jonathan Cornelissen is one of the co-founders of DataCamp, and is interested in everything related to data science, R, education and entrepreneurship. He holds a PhD in financial econometrics, and is the author of an R package for quantitative finance. DataCamp is the second education start-up he founded, and the first one that went international. In

Obsess about R

http://www.r-bloggers.com/

R-bloggers
R news and tutorials contributed by (573) R bloggers

Home About RSS add your blog! R jobs Contact us

WELCOME!

Here you will find daily news and tutorials about R, contributed by over 573 bloggers. There are many ways to follow us - By e-mail:

On Facebook:

If you are an R blogger yourself you are invited to add your own R content feed to this site (Non-English R bloggers should add themselves- here)

JOBS FOR R-USERS
Quantitative Developer (@New York)

functional enrichment analysis with NGS data

August 20, 2015
By ygc

I found that there is a Bioconductor package, seq2pathway, that can apply functional analysis to NGS data. It consists of two components, seq2gene and gene2pathway. seq2gene converts genomic coordination to genes while gene2pathway performs functional analysis at gene level. Read More: 1007 Words Totally

[Read more »](#)

Deploying a car price model using R and AzureML

August 20, 2015
By Longhow Lam

Data Manipulation with dplyr

August 20, 2015
By Teja Kodali

Recently Microsoft released the AzureML R package, it allows R users to publish their R models (or any R function) as a web service on

TOP 3 POSTS FROM THE PAST 2 DAYS

- 5 New R Packages for Data Scientists
- In-depth introduction to machine learning in 15 hours of expert videos
- Installing R packages

Search & Hit Enter

TOP 9 ARTICLES OF THE WEEK

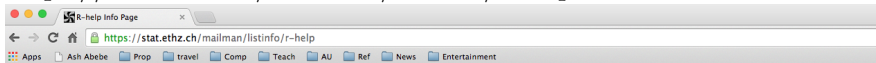
1. In-depth introduction to machine learning in 15 hours of expert videos
2. Scatterplots
3. Installing R packages
4. Using apply, sapply, lapply in R
5. Playing with R, Shiny Dashboard and Google Analytics Data
6. Importing Data Into R - Part Two
7. Yet another post on google scholar data analysis
8. Basics of Histograms
9. R 3.2.2 is released

SPONSORS

EARI

Ask questions about R

https://stat.ethz.ch/mailman/listinfo/r-help



R-help -- Main R Mailing List: Primary help

About R-help

The main R mailing list, for announcements about the development of R and the availability of new code, questions and answers about problems and solutions using R, enhancements and patches with S and S-plus, and for the posting of nice examples and benchmarks. Please read the [General Instructions](#) on the [R Mailing Lists](#) page and follow the [posting guide](#)!

This has become quite an active list with often dozens of messages per day. It might make sense therefore to check the *digest* box in your subscription page.

For discussion about new features or R's future development, use the [R-devel](#) mailing list. See the [R Project](#) site for more information about R. Before asking, please read [the FAQ](#) about R.

To see the collection of prior postings to the list, visit the [R-help Archives](#) --- or use the [searchable archives](#) provided by Robert King and U. Newcastle.AU or the [full R site search](#) provided by Jc interfaces.

Posters should be aware that the R lists are *public* discussion lists and anything you post will be **archived and accessible** via several websites for many years.

Using R-help

Since Jan.31, 2008, to be allowed to freely post messages, you (i.e., your sending e-mail address) **must be subscribed** to the list.

Thanks to a dozen of **volunteer** moderators, posting of non-subscribers is still possible -- with a delay because of the moderator approval needed. The volunteers, generously devoting part of their Robin Hankin, Klaus Nordhausen, Anne York, Adrian Dusa, Kevin Thorpe, Peter Alspach and David Winsemius (in addition to the R-help list maintainers).

To post a message to all the list members, send email to r-help@r-project.org.

You can subscribe to the list, or change your existing subscription, in the sections below.

Subscribing to R-help

Subscribe to R-help by filling out the following form. You will be sent email requesting confirmation, to prevent others from gratuitously subscribing you. This is a hidden list, which means that it

Your email address:

Your name (optional):

You may enter a privacy password below. This provides only mild security, but should prevent others from messing with your subscription. **Do not use a valuable password** as it will occasionally be emailed back to you in cleartext.

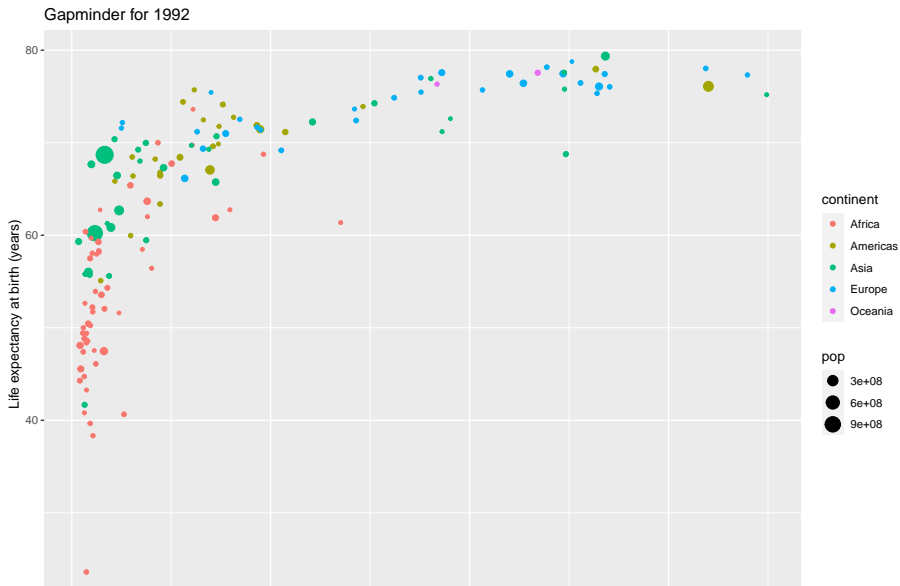
Your Turn :Use Gapminder data

```
library(gapminder)
library(dplyr)
df <- gapminder::gapminder
df1 = df %>% filter(year == 1992)
df1
```

```
## # A tibble: 142 x 6
```

##	country	continent	year	lifeExp	pop	gdpPercap
##	<fct>	<fct>	<int>	<dbl>	<int>	<dbl>
## 1	Afghanistan	Asia	1992	41.7	16317921	649.
## 2	Albania	Europe	1992	71.6	3326498	2497.
## 3	Algeria	Africa	1992	67.7	26298373	5023.
## 4	Angola	Africa	1992	40.6	8735988	2628.
## 5	Argentina	Americas	1992	71.9	33958947	9308.
## 6	Australia	Oceania	1992	77.6	17481977	23425.
## 7	Austria	Europe	1992	76.0	7914969	27042.
## 8	Bahrain	Asia	1992	72.6	529491	19036.

Produce this Plot



THANK YOU

- **Thank You ALL**