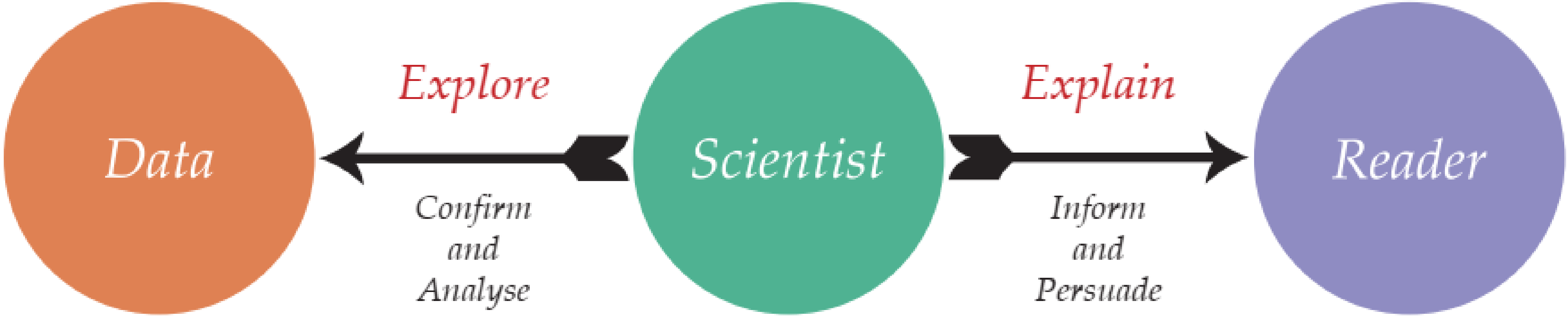
DATA
SCIENCE
INDONESIA

Data Visualization

Universitas Airlangga, 28 Agustus 2018

Achmad Wildan Al Aziz
Fiika Arma’atus Syaani

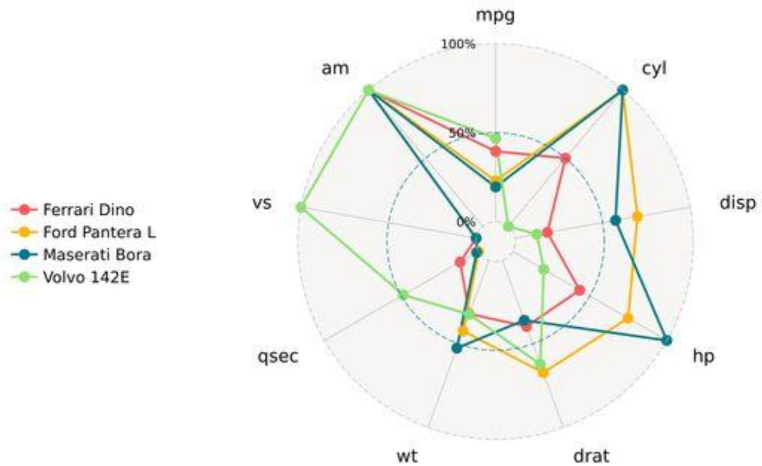

Data
Explore
Confirm and Analyse
Scientist
Explain
Inform and Persuade
Reader

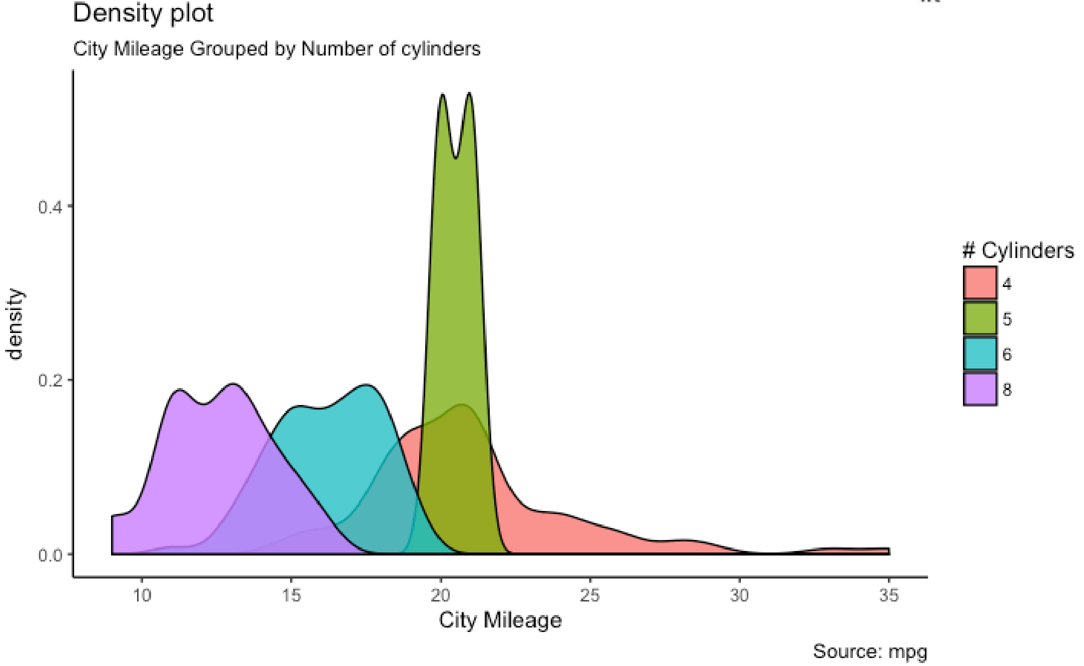
Outline

DATA SCIENCE INDONESIA

- Base Graphics vs ggplot2
- The Grammar Of Graphics
- Geometric Objects And Aesthetics
- Points
- Statistical Transformation
- Scales

Example

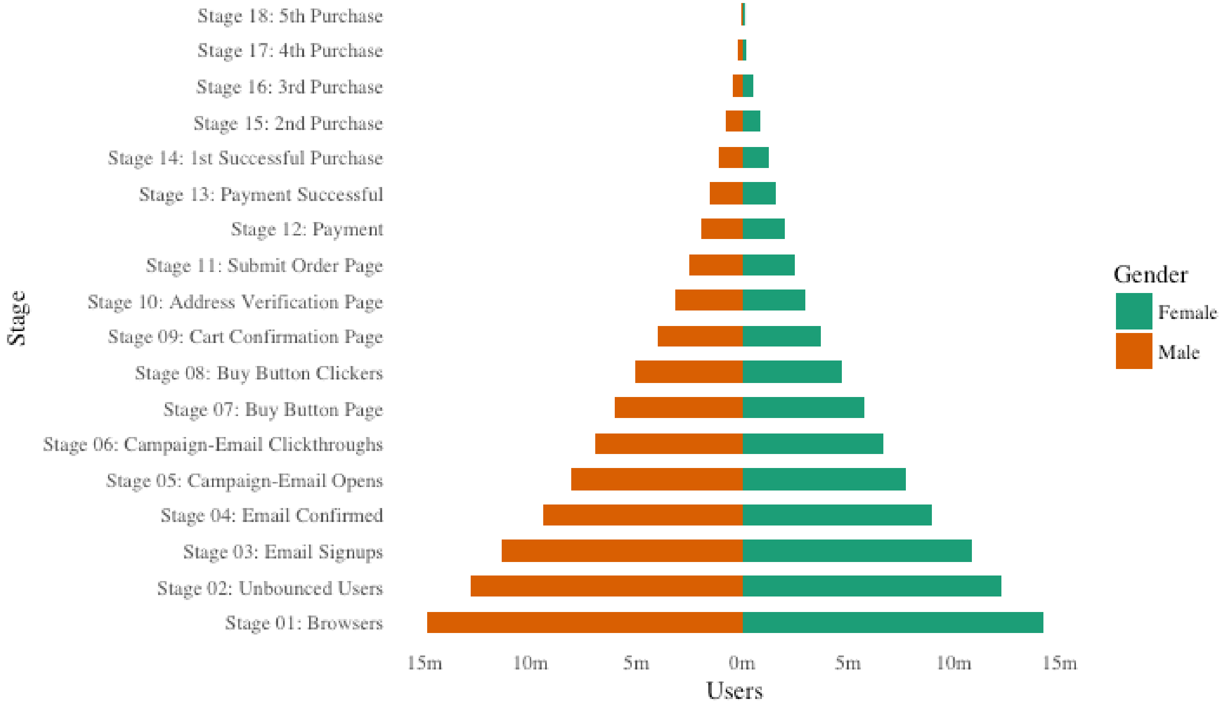



Density plot

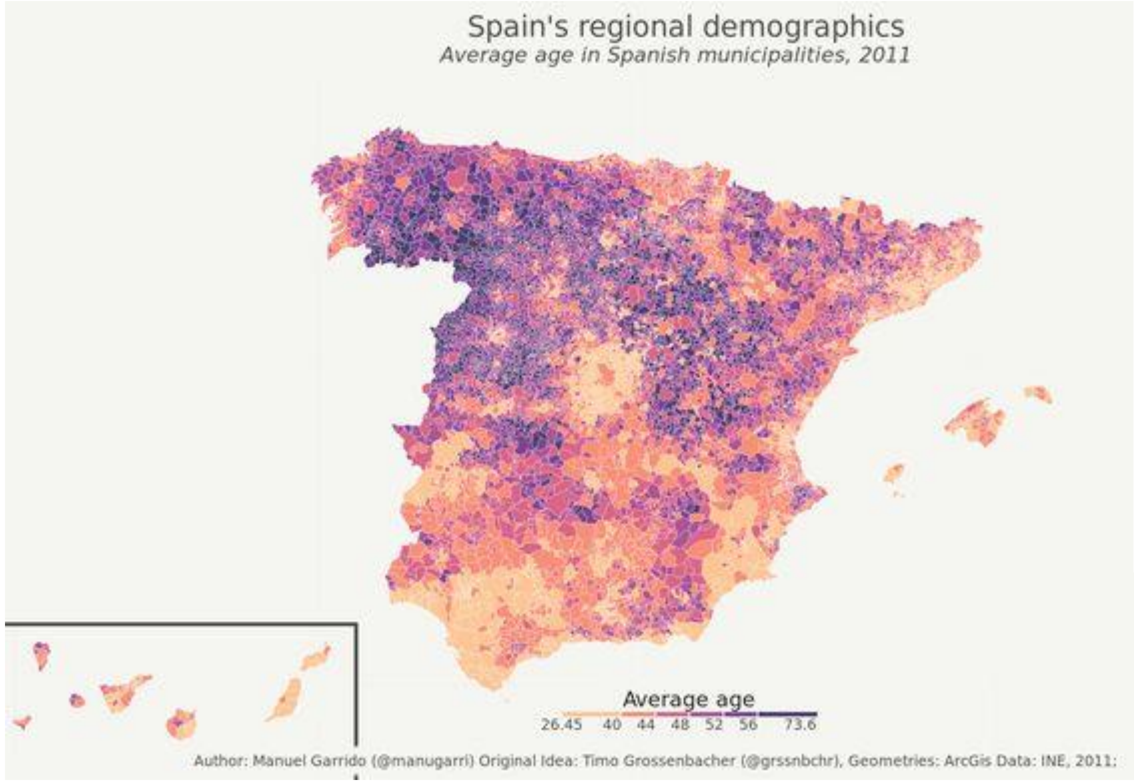
City Mileage Grouped by Number of cylinders

Source: mpg

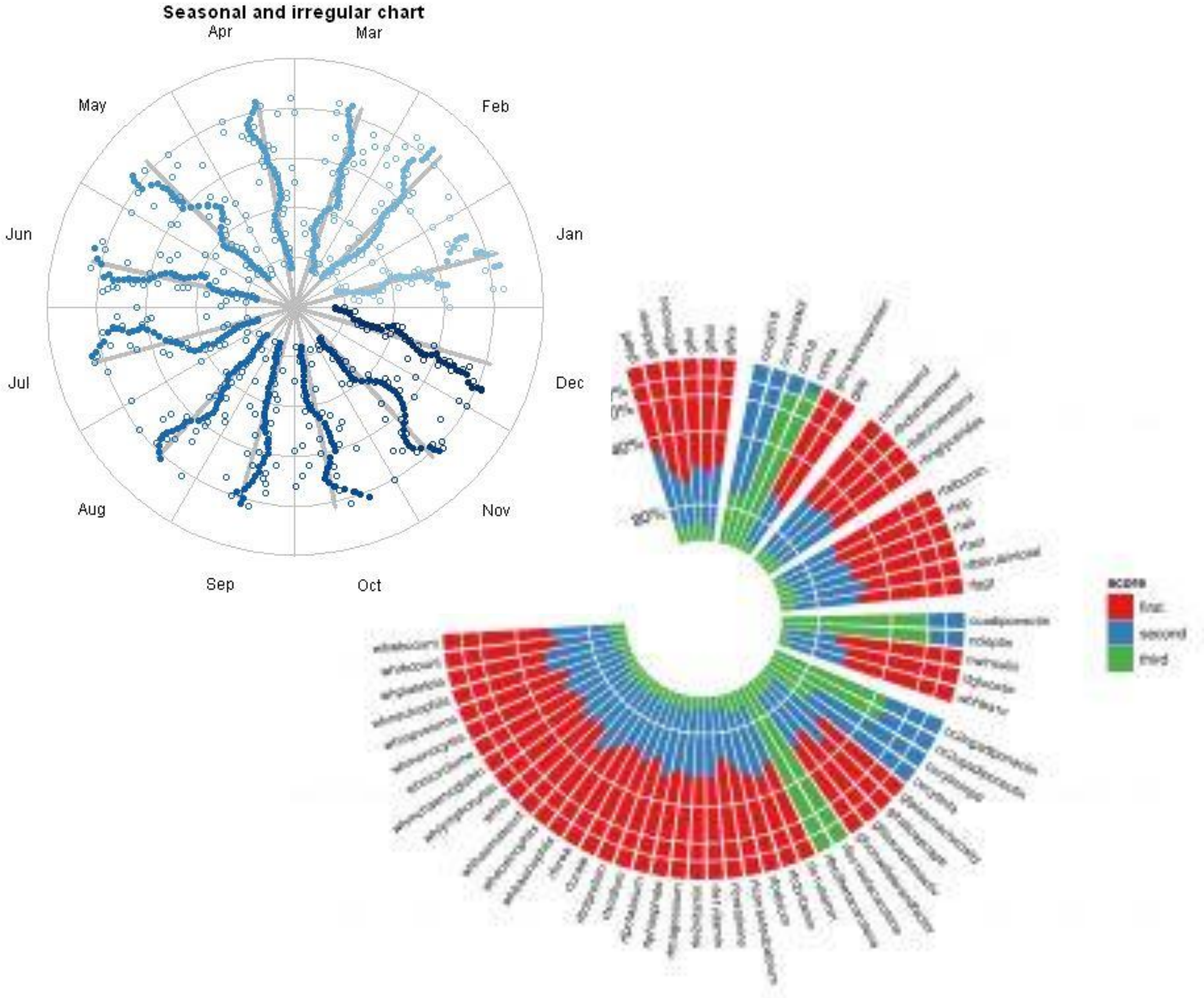
Email Campaign Funnel



Example


Spain's regional demographics
Average age in Spanish municipalities, 2011
Average age
26.45
40
44
48
52
56
73.6

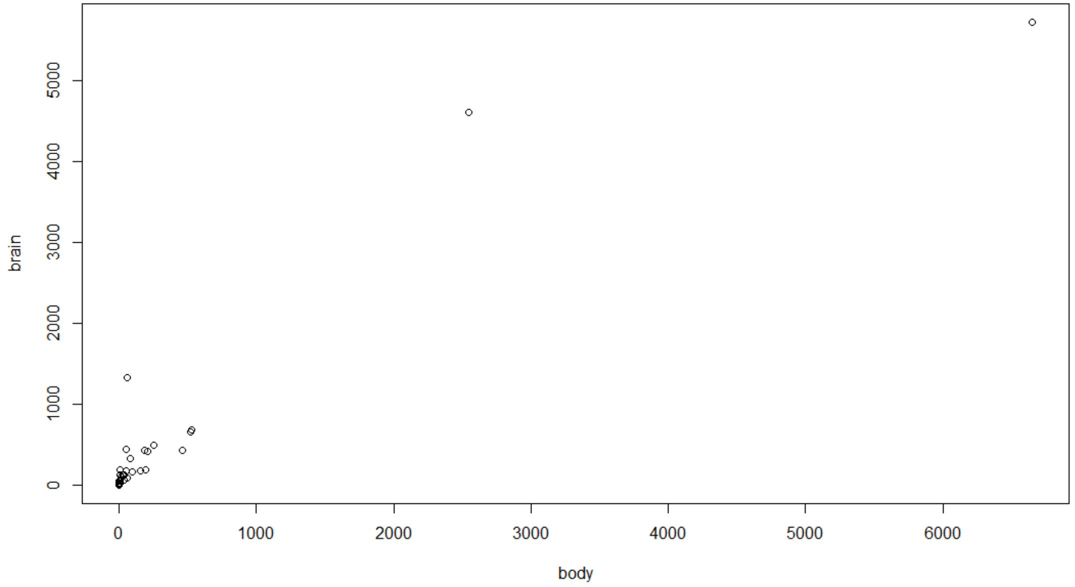
Author: Manuel Garrido (@manugarri) Original Idea: Timo Grossenbacher (@grssnbchr), Geometries: ArcGis Data: INE, 2011;

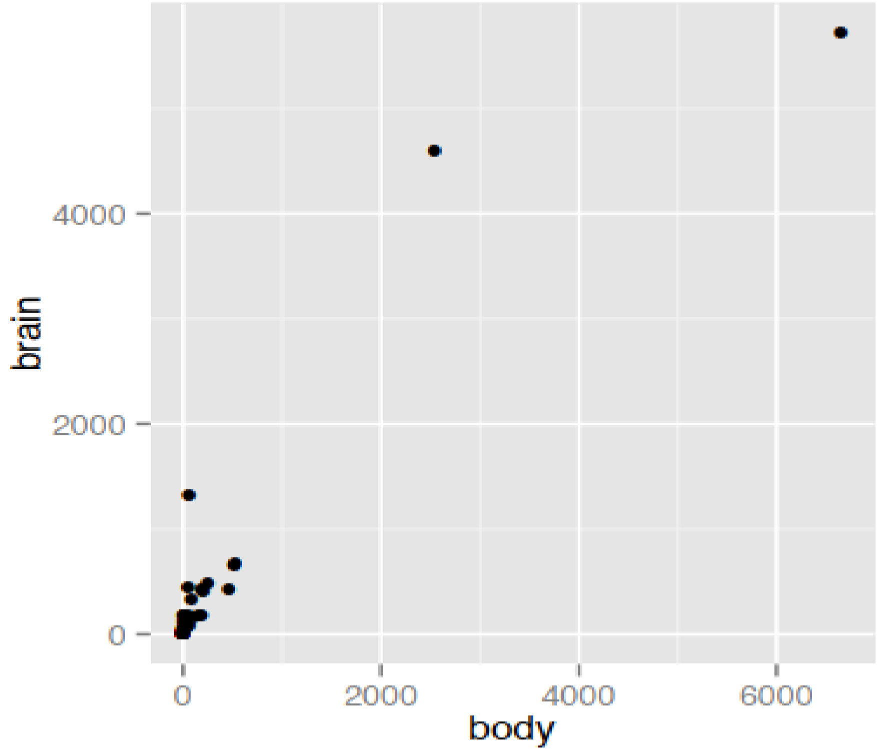


Base Graphics vs ggplot2

```
> plot(mammals)
```

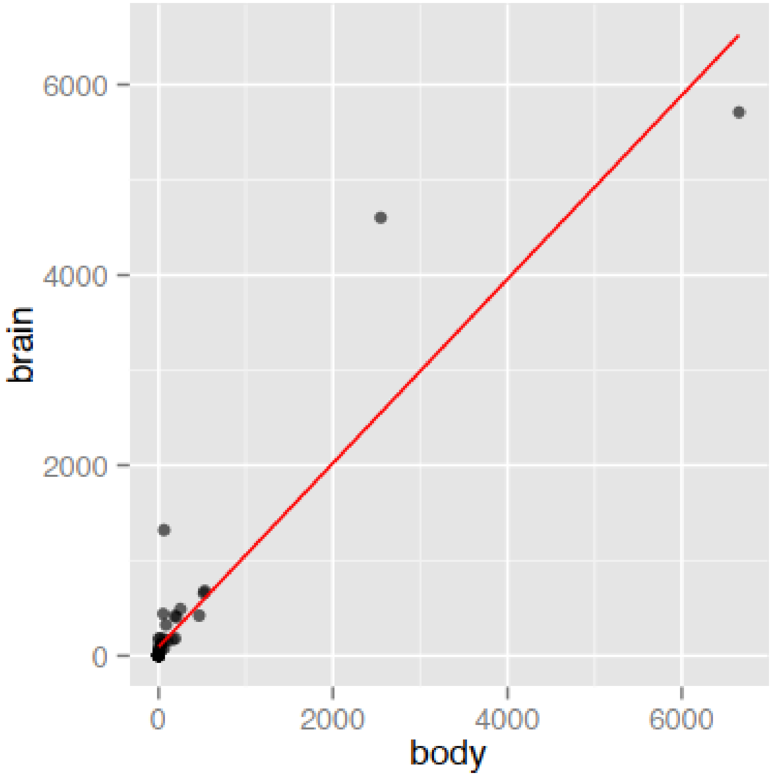
```
> ggplot(mammals, aes(x =
  body, y = brain)) +
  geom point()
```





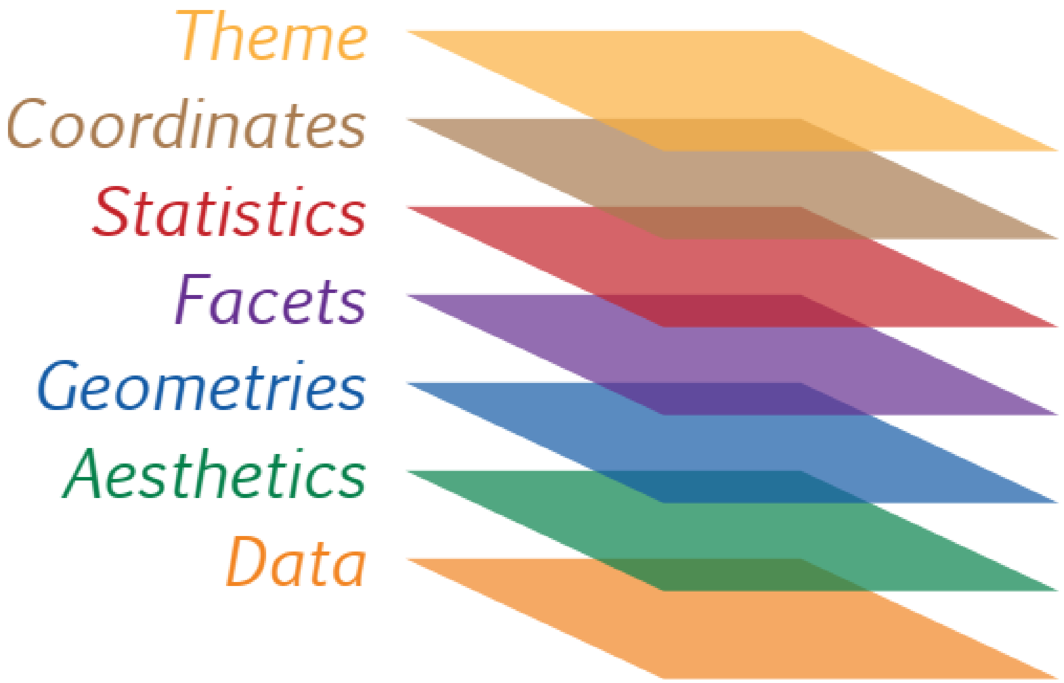
Base Graphics vs ggplot2

```
> ggplot(mammals, aes(x = body, y = brain)) +
      geom_point(alpha = 0.6) +
      stat_smooth(method = "lm", col = "red", se = FALSE)
```



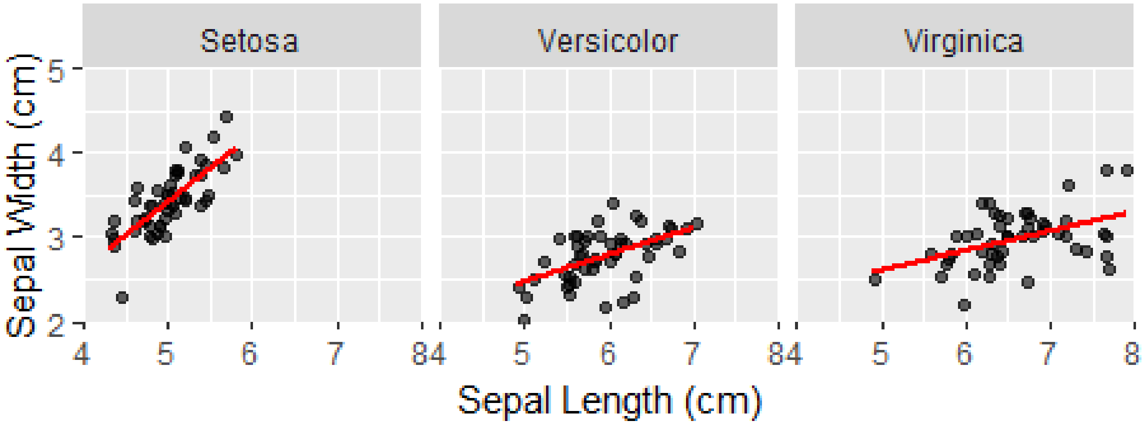
The Grammar Of Graphics

- data
- aesthetic mapping
- geometric object
- statistical transformations
- scales
- coordinate system
- position adjustments
- faceting



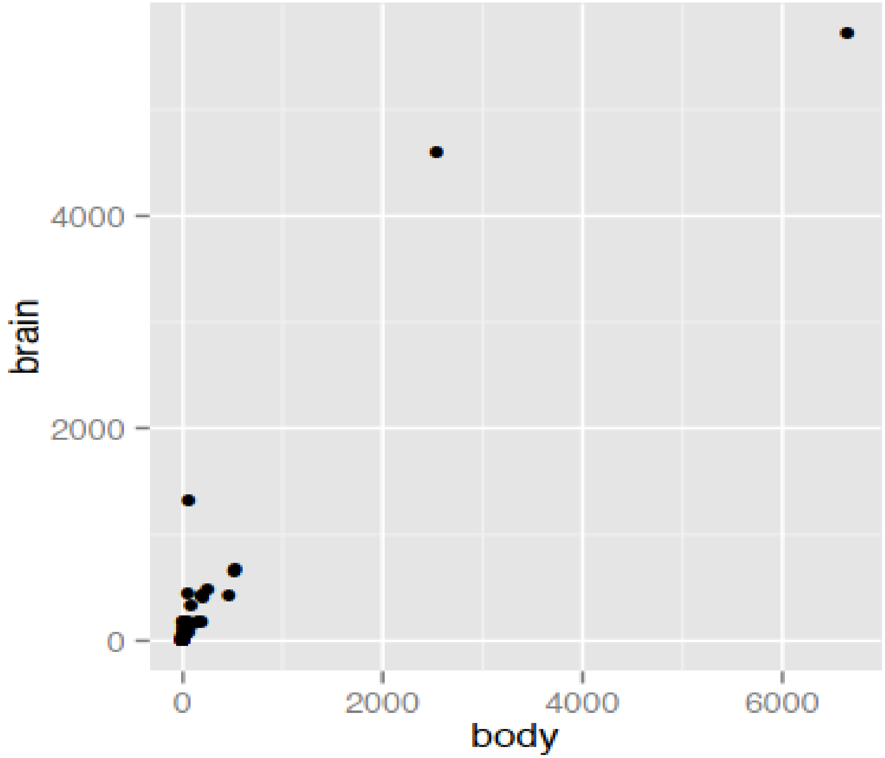
The Grammar Of Graphics

```
> ggplot(iris, aes(x = Sepal.Length, y = Sepal.Width)) +
     geom_jitter(alpha = 0.6) +
     facet_grid(. ~ Species) +
     stat_smooth(method = "lm", se = F, col = "red") +
     scale_y_continuous("Sepal Width (cm)",
             limits = c(2,5), expand = c(0,0)) +
     scale_x_continuous("Sepal Length (cm)",
             limits = c(4,8), expand = c(0,0)) +
     coord_equal()
```



Geometric Objects And Aesthetics

```
> ggplot(mammals, aes(x = body, y = brain)) +
  geom_point()
```



Geometric Objects And Aesthetics

Aesthetic Mapping (`aes`)

aesthetic properties to represent variables and "something you can see"

- position (i.e., on the x and y axes)
- color ("outside" color)
- fill ("inside" color)
- shape (of points)
- alpha (transparency)
- linetype
- labels (text on a plot or axes)
- size

Geometic Objects (`geom`)

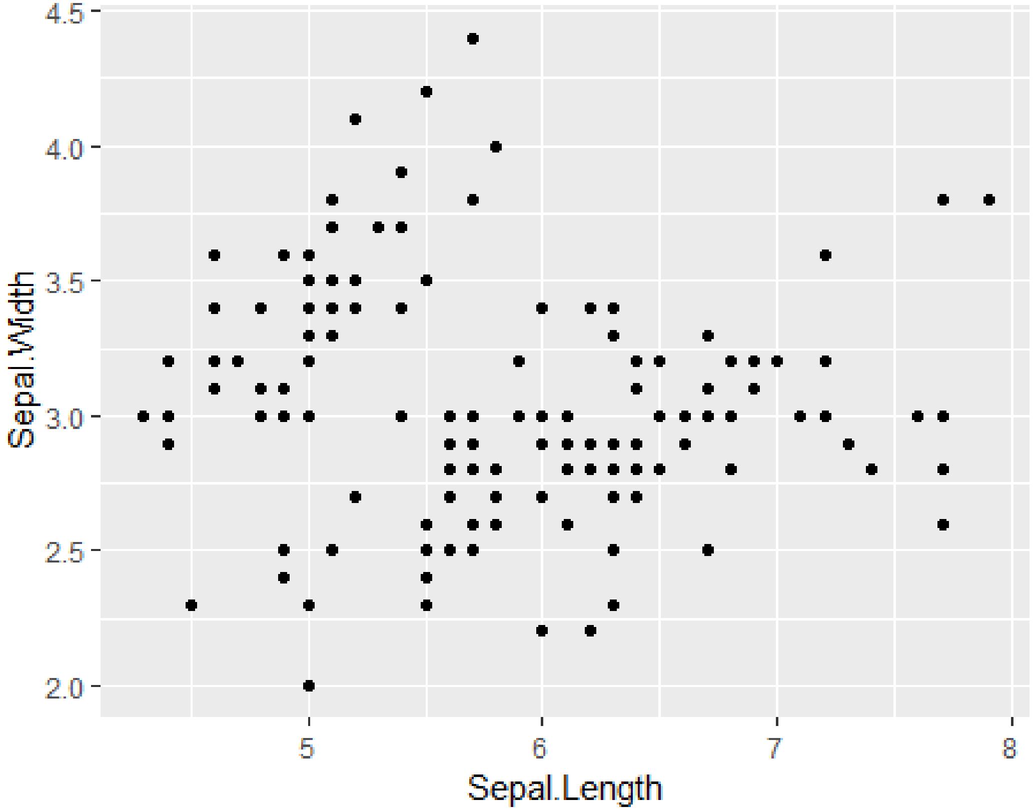
Use a geom function to represent data points.

A plot must have at least one **geom**; there is no upper limit. You can add a geom to a plot using the + operator

- points (`geom_point`, for scatter plots, dot plots, etc)
- lines (`geom_line`, for time series, trend lines, etc)
- boxplot (`geom_boxplot`)
- …

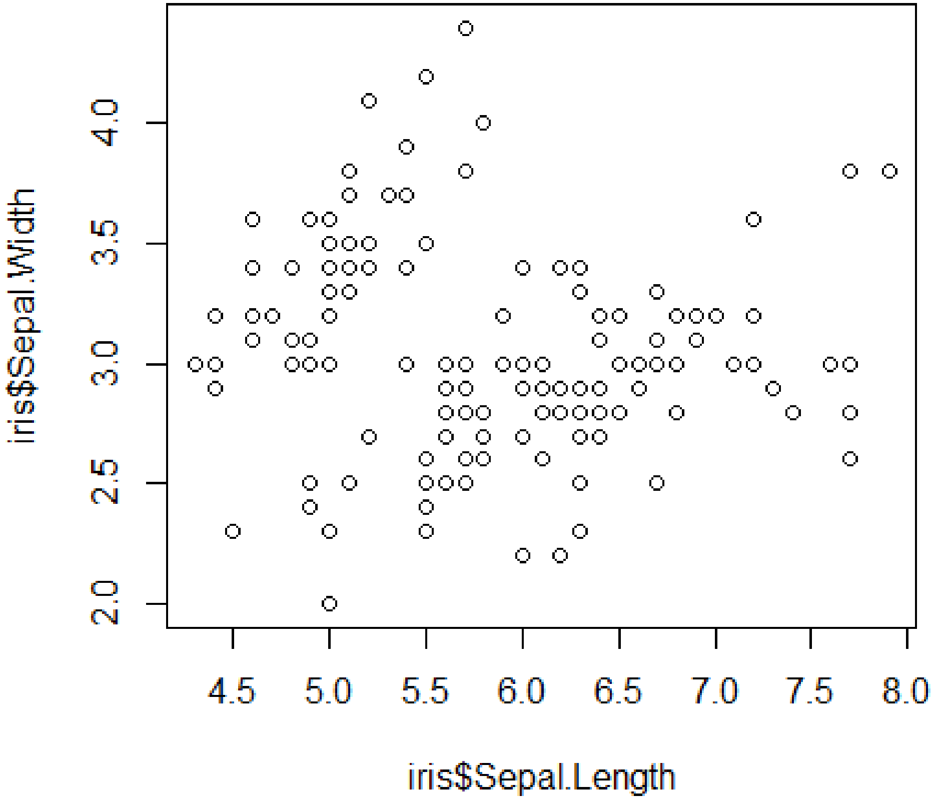
Points (geom_point)

```
> ggplot(iris, aes(x = Sepal.Length, y = Sepal.Width)) +
      geom_point()
```



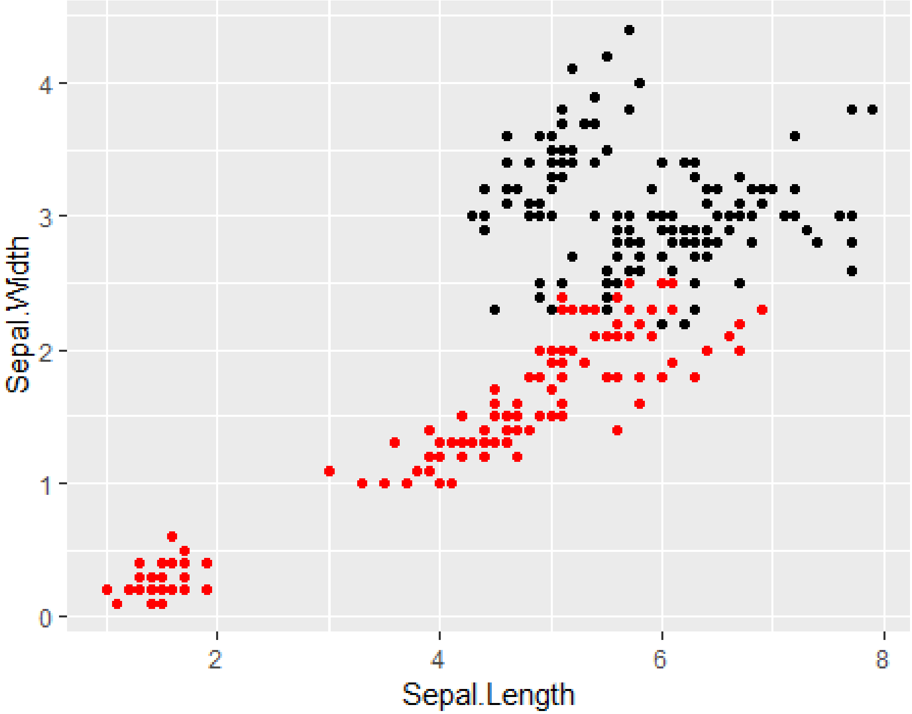
Points (geom_point)

```
> plot(iris$Sepal.Length, iris$Sepal.Width)
```



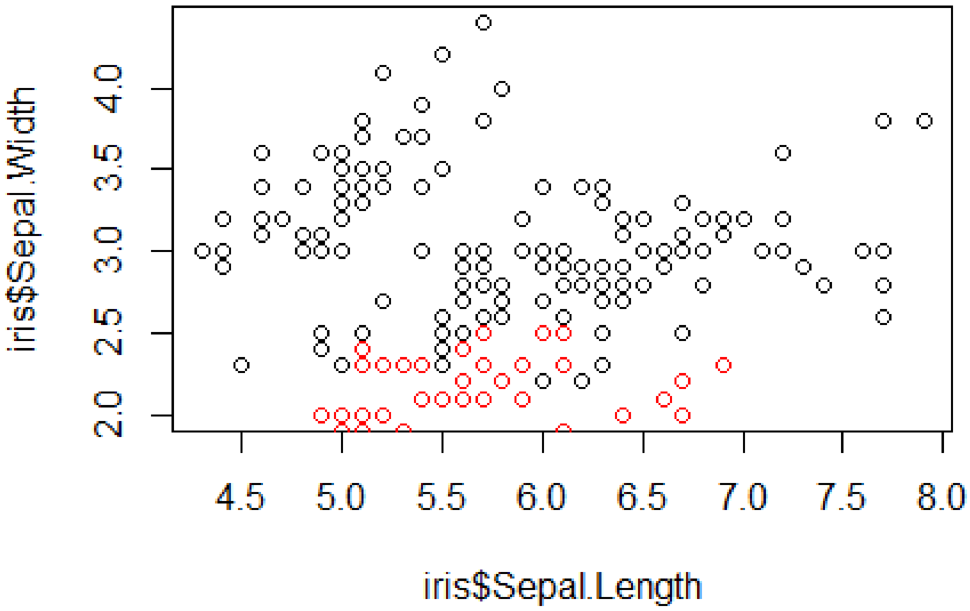
Points (geom_point)

```
> ggplot(iris, aes(x = Sepal.Length, y = Sepal.Width)) +
      geom_point() +
      geom_point(aes(x = Petal.Length, y = Petal.Width), col =
            "red")
```



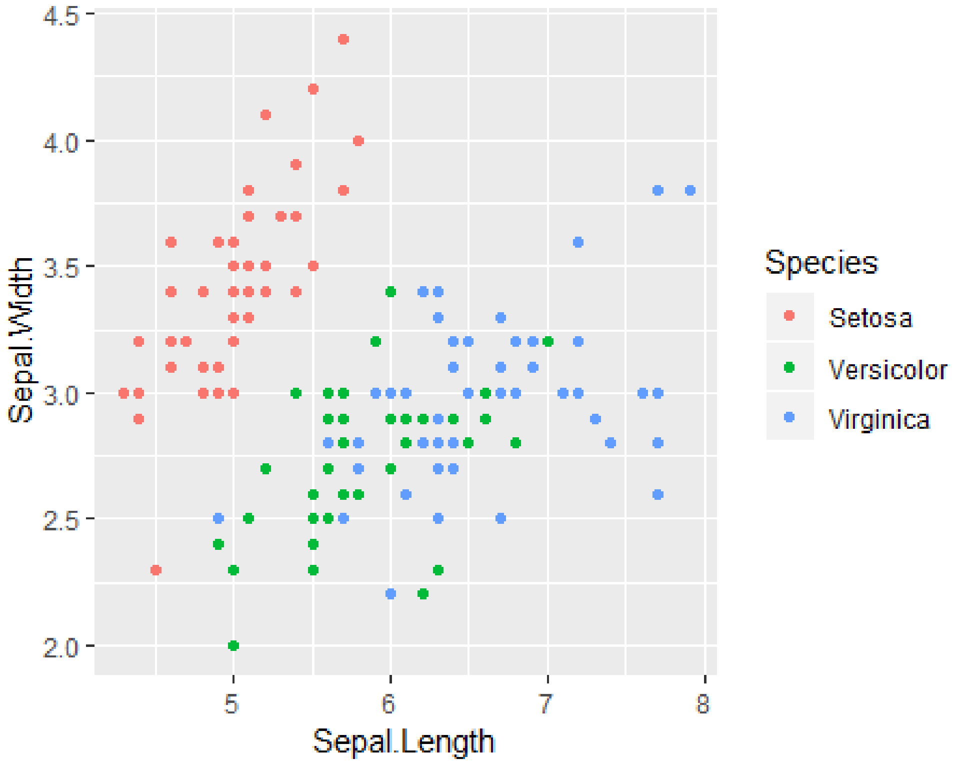
Points (geom_point)

```
> plot(iris$Sepal.Length, iris$Sepal.Width)
> points(iris$Petal.Length, iris$Petal.Width, col = "red")
```



Points (`geom_point`)

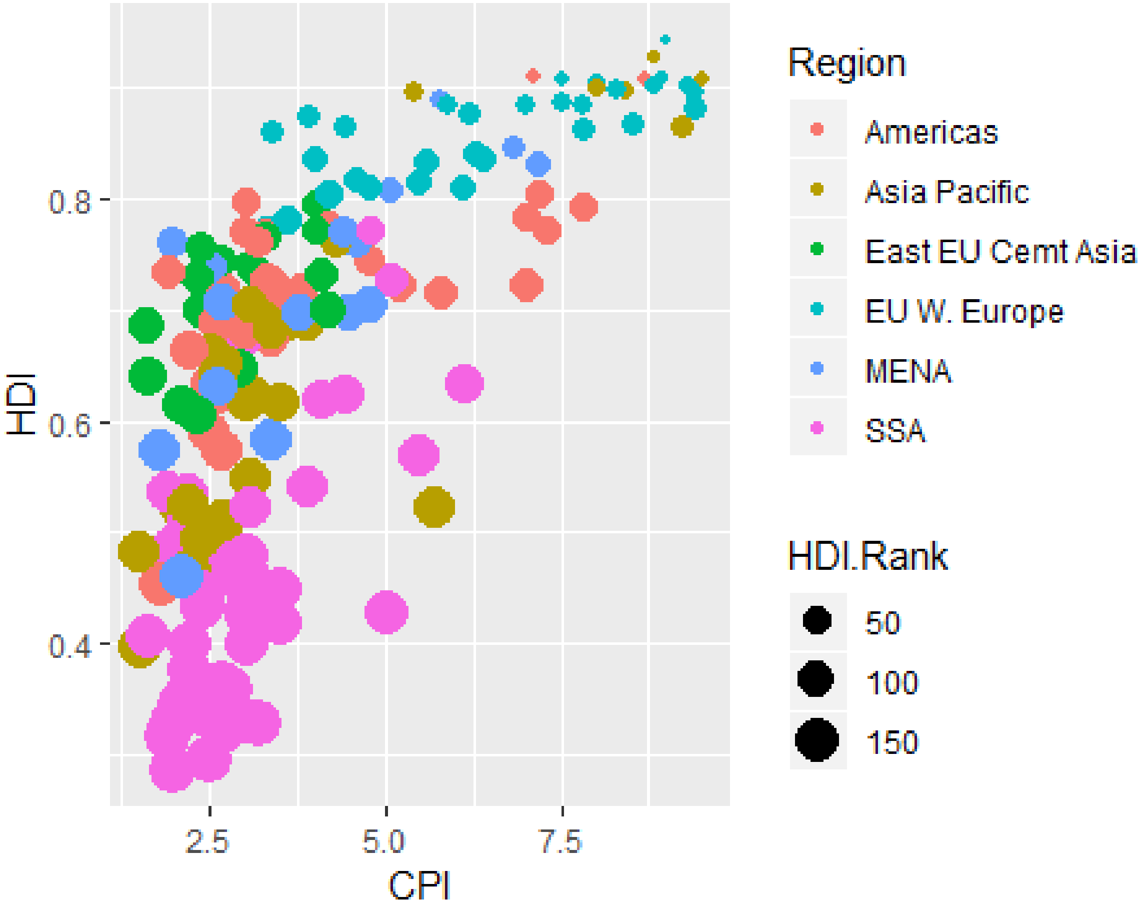
```
> ggplot(iris) +
   geom_point(aes(x = Sepal.Length, y = Sepal.Width,col = Species))
```



Exercise 1

Data = EconomistData.csv

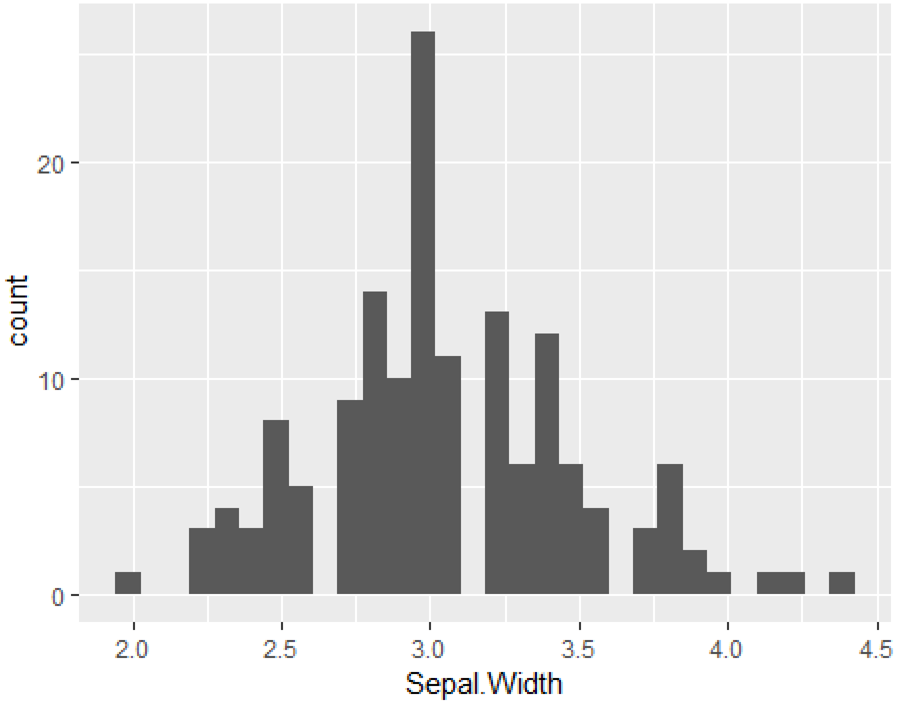
1. Create a scatter plot with CPI on the x axis and HDI on the y axis.
2. Color the points blue.
3. Map the color of the points to Region.
4. Make the points bigger by setting size to 2
5. Map the size of the points to HDI.Rank



Statistical Transformation (stat_)

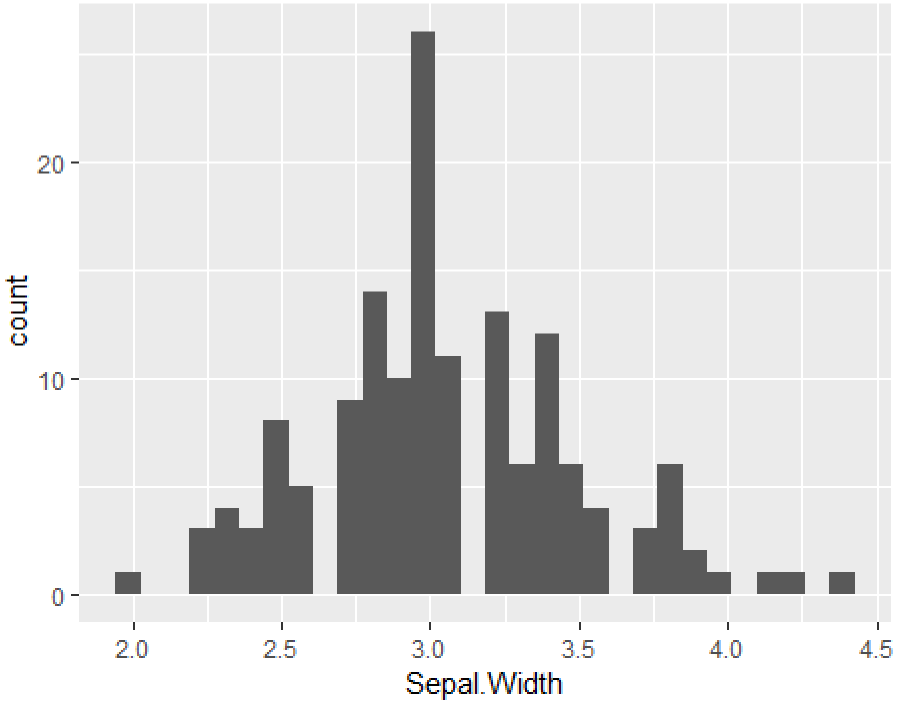
Called from within a geom

```
> ggplot(iris, aes(x =
      Sepal.Width)) +
      geom_histogram()
```



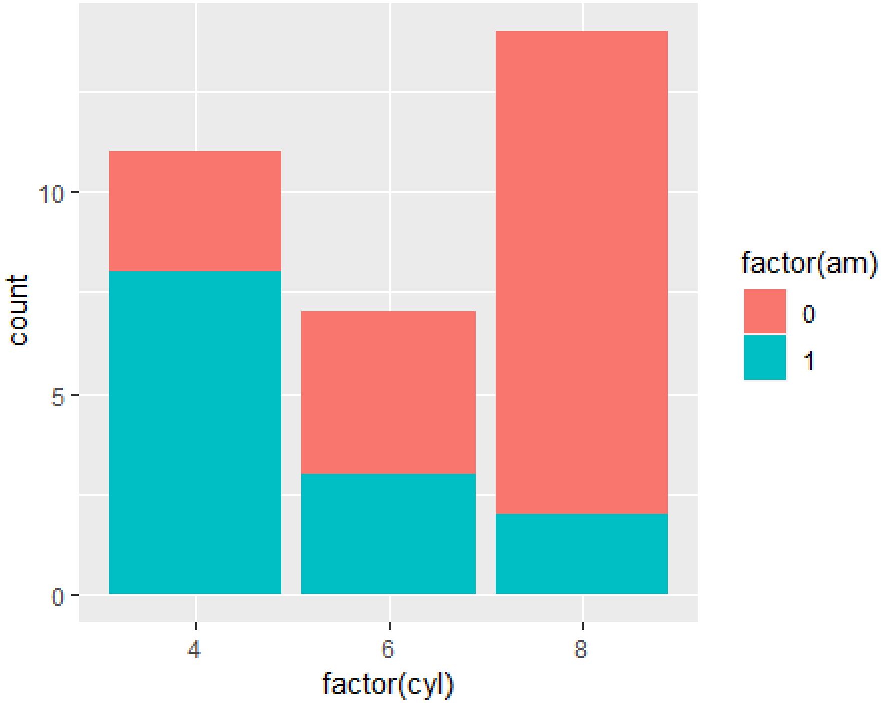
Called independently with stat_

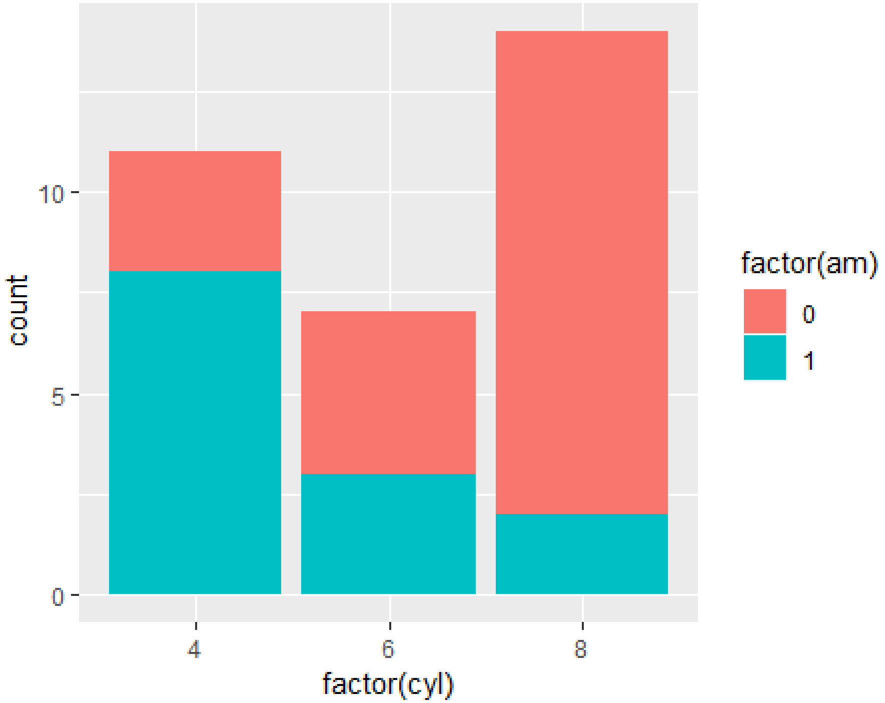
```
> ggplot(iris, aes(x =
      Sepal.Width)) +
      stat_bin()
```



Statistical Transformation (`stat_`)

```
> ggplot(mtcars, aes(x = factor(cyl), fill = factor(am)) +
      geom_bar()
> ggplot(mtcars, aes(x = factor(cyl), fill = factor(am)) +
       stat_bin()
```



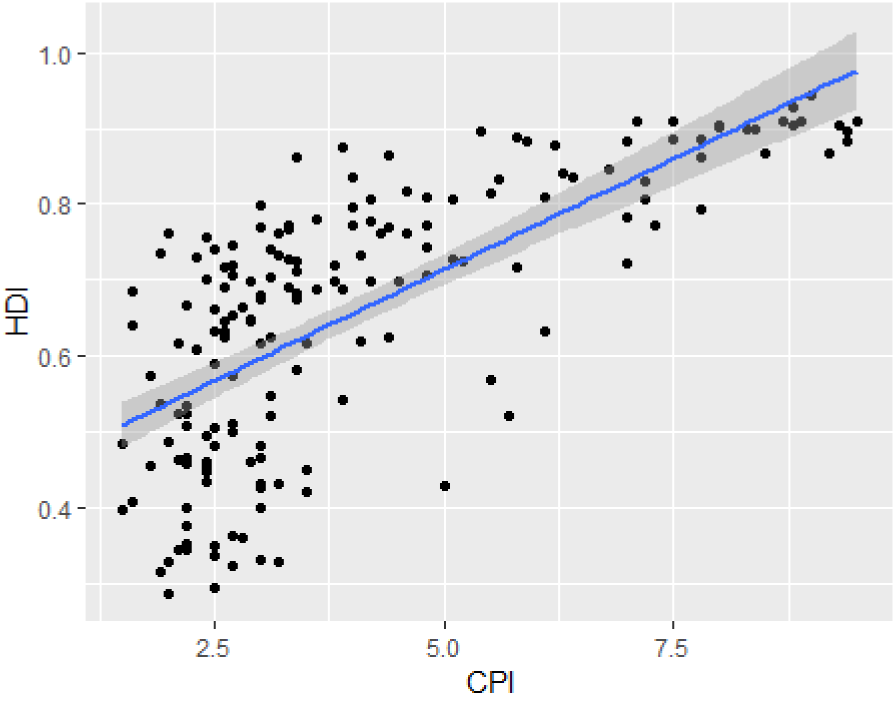


Statistical Transformation (stat_)

stat_	geom_
stat_bin()	geom_histogram()
stat_bin()	geom_bar()
stat_bin()	geom_freqpoly()
stat_smooth()	geom_smooth()
stat_boxplot()	geom_boxplot()
stat_bindot()	geom_dotplot()
stat_bin2d()	geom_bin2d()
stat_binhex()	geom_hex()
stat_contour()	geom_contour()
stat_quantile()	geom_quantile()
stat_sum()	geom_count()

Exercise 2

1. Re-create a scatter plot with CPI on the x axis and HDI on the y axis (as you did in the previous exercise).
2. Overlay a smoothing line on top of the scatter plot using geom_smooth / stat_smooth .
3. Overlay a smoothing line on top of the scatter plot using geom_smooth / stat_smooth, but use a linear model for the predictions.

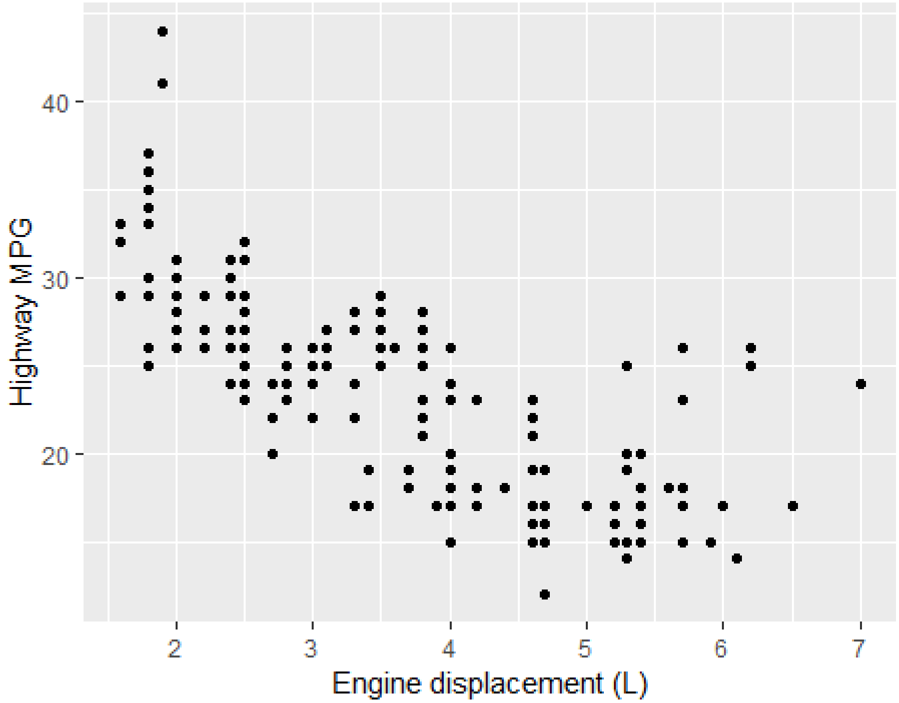


Scales

map data values to the visual values of an aesthetic.

Scales are reported on the plot using axes and legends

```
> ggplot(mpg, aes(displ, hwy)) +
   geom_point() +
   scale_x_continuous("Engine displacement (L)")+
   scale_y_continuous("Highway MPG")
```

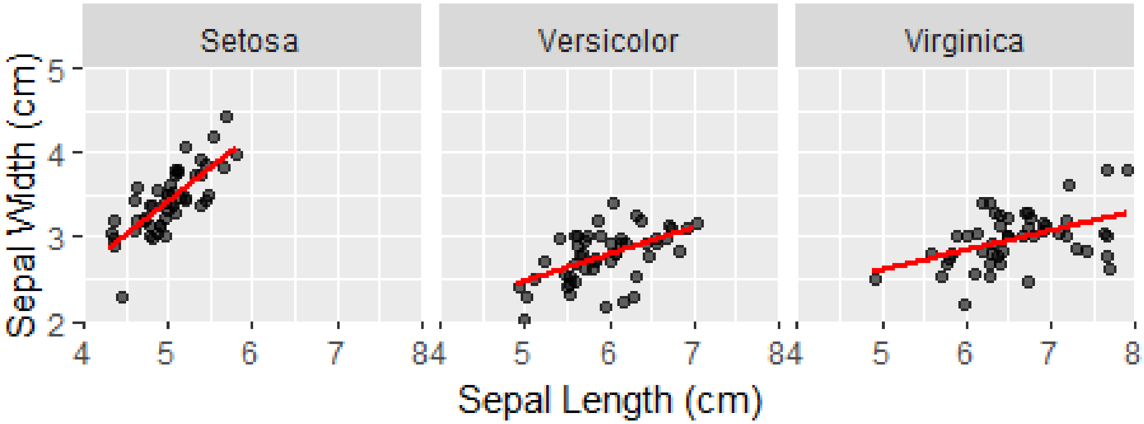


Exercise 3

1. Create a scatter plot with CPI on the x axis and HDI on the y axis. Color the points to indicate region.
2. Modify the x, y, and color scales so that they have more easily-understood names (e.g., spell out “Human development Index” instead of “HDI”).

Putting it All Together

```
> ggplot(iris, aes(x = Sepal.Length, y = Sepal.Width)) +
    geom_jitter(alpha = 0.6) +
    facet_grid(. ~ Species) +
    stat_smooth(method = "lm", se = F, col = "red") +
    scale_y_continuous("Sepal Width (cm)",
              limits = c(2,5), expand = c(0,0)) +
    scale_x_continuous("Sepal Length (cm)",
              limits = c(4,8), expand = c(0,0)) +
    coord_equal()
```


Setosa
Versicolor
Virginica
Sepal Width (cm)
Sepal Length (cm)