Data visualization with ggplot2

* Merge statistics with design in meaningful and appropriate ways
  + Graphical data analysis emphasizing accurate rep and interp of data
  + Attractive and meaningful – aids understanding and comm of results
* Exploratory vs explanatory
  + Exploratory – easily generated, data heavy, intended for small specialist audience, confirm and analyze
  + Explanatory – labor intensive, data specific, intended for broad audience like publication, communications, inform and persuade
* Good design begins with thinking about the audience
* MASS::mammals

library(MASS)

mammals

library(ggplot2)

# explore – scatter plot

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

geom\_point() +

# explore – linear model

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

geom\_point(alpha = 0.6)

stat\_smooth(method = “lm”, col = “red”, se = FALSE)

# explore – fine tuning axes

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

geom\_point(alpha = 0.6) +

coord\_fixed() +

scale\_x\_log10() +

scale\_y\_log10() +

stat\_smooth(method = “lm”, col = “#C42126”, se = FALSE, size = 1)

# Publication ready plot

library(scales) # functions trans\_breaks and trans\_format

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

annotation\_logticks() +

geom\_point(alpha = 0.6) +

coord\_fixed(xlim = c(10^-3, 10^4), ylim = c(10^-1, 10^4)) +

scale\_x\_log10(expression(“Body weight (log”[10”]\*”(Kg))”),

breaks = trans\_breaks(“log10”, function(x) 10^x),

labels = trans\_format(“log10”, math\_format(10^.x))) +

scale\_y\_log10(expression(“Brain weight (log”[“10”]\* “(g))”),

breaks = trans\_breaks(“log10”, function(x) 10^x),

labels = trans\_format(“log10”, math\_format(10^.x))) +

stat\_smooth(method = “lm”, col = “#C42126”, se = FALSE, size = 1) +

theme\_classic

* Usefulness of data visualization as data analysis trend – Anscombe plots
  + Consider same lm that could be describing a real good fit relationship, a parabolic relationship, a lm that should be diff with an outlier, and no relationship but an outlier that drives an apparent relationship through the viewed mathematical model only
  + If you only model and don’t visualize data, you miss those trends which are all distinct and interesting
* Some amount of trial and error in a creative process
* Understand how structure of data set can help determine useful types of plots
* How to use elements such as color and size effectively
* Best plot type for accurately representing nature of data
* Common pitfalls and how to avoid them
* Course 1: concepts, data, aesthetics, geometries
* Course 2: statistics, coordinates, facets, themes, best practices, case study
* Course 3: Advanced plots and ggplot2 internals
* Grammar of graphics
* Plotting framework
* 2 principles
* Graphics = distinct layers of grammatical elements
* Meaningful plots through aesthetic mappings – grammatical rules for assembling plots
* Grammatical elements
  + \*Data – dataset being plotted
  + \*Aesthetics – scales onto which we map our data
    - X-axis
    - Y-axis
    - Color
    - Fill
    - Size
    - Labels
    - Alpha
    - Shape
    - Line width
    - Line type
  + \*Geometries – visual elements used for our data
    - Point
    - Line
    - Histogram
    - Bar
    - boxplot
  + Coordinates – space on which data will be plotted
    - Cartesian
    - Fixed
    - Polar
    - limits
  + Statistics – representations of our data to aid understanding
    - Binning
    - Smoothing
    - Descriptive
    - inferential
  + Facets – plotting small multiples
    - Rows
    - Columns
  + Themes – all non-data ink
* Iris data set – collected in 1930s by Edgar Anderson and popularized by R.A. Fischer
  + 3 types of iris: setosa, versicolor, virginica
  + 4 measurements: petal length, petal width, sepal length, sepal width
  + 50 of each species

levels(iris$Species) <- c(“Setosa”, “Versicolor”, “Virginica”)

ggplot(iris, aes(x = Sepal.Length, y = Sepal.Width)) + # data, aesthetics layers

geom\_jitter(alpha = 0.6) + # geometries layer

facet\_grid(. ~ Species) + # facets layer

stat\_smooth(method = “lm”, se = FALSE, col = “red”) + # statistics layer

scale\_y\_continuous(“Sepal Width (cm)”, limits = c(2, 5), expand = c(0, 0)) + # coordinates

scale\_x\_continuous(“Sepal Length (cm)”, limits = c(4, 8), expand = c(0, 0)) + # coordinates

coord\_equal() +

theme(panel.background = element\_blank(),

plot.background = element\_blank(),

legend.background = element\_blank(),

legend.key = element\_blank(),

strip.background = element\_blank(),

axis.text = element\_text(colour = “black”),

axis.ticks = element\_line(colour = “black”),

panel.grid.major = element\_blank(),

panel.grid.minor = element\_blank(),

axis.line = element\_line(colour = “black”),

strip.text = element\_blank(),

panel.margin = unit(1, “lines”)

)