Models

INFO 201

Today’s Objectives

Highlights from "the industry"

Introduce linear modeling

Explore packages for advanced data management

Plotcon

Safia Abdalla

ABOUT PUBLIC SPEAKING CONSULTING BLOG

I'm a data scientist and software engineer with an interest in open source software and data science for social good.

`x`

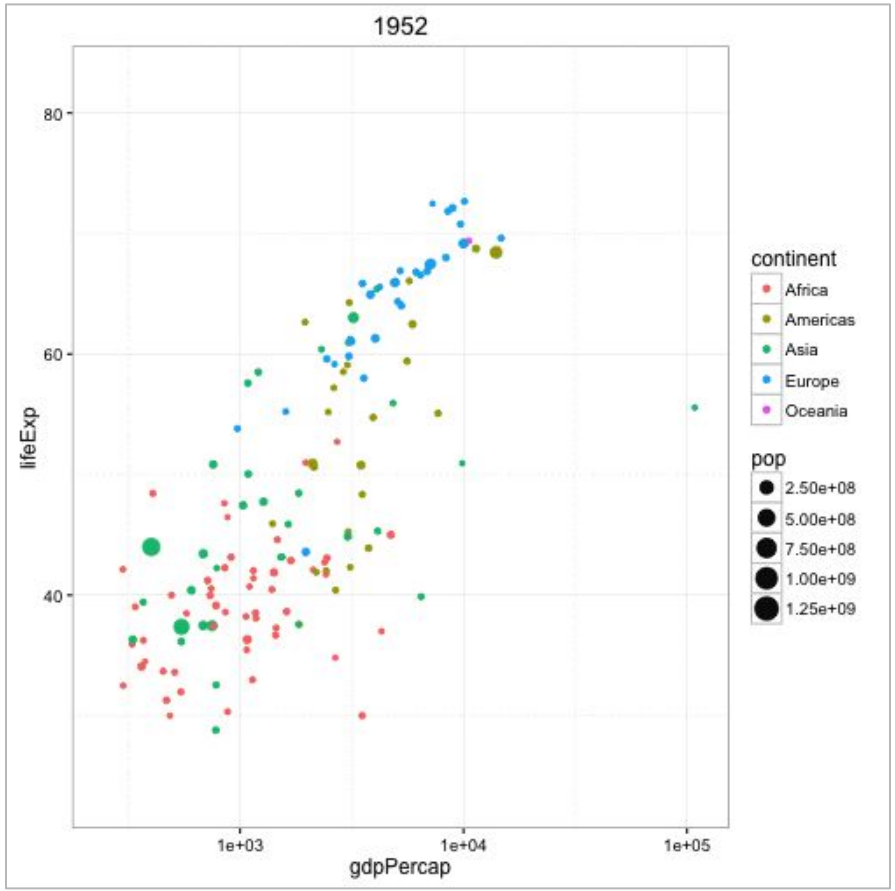
`x[i]`

`x[[i]]`

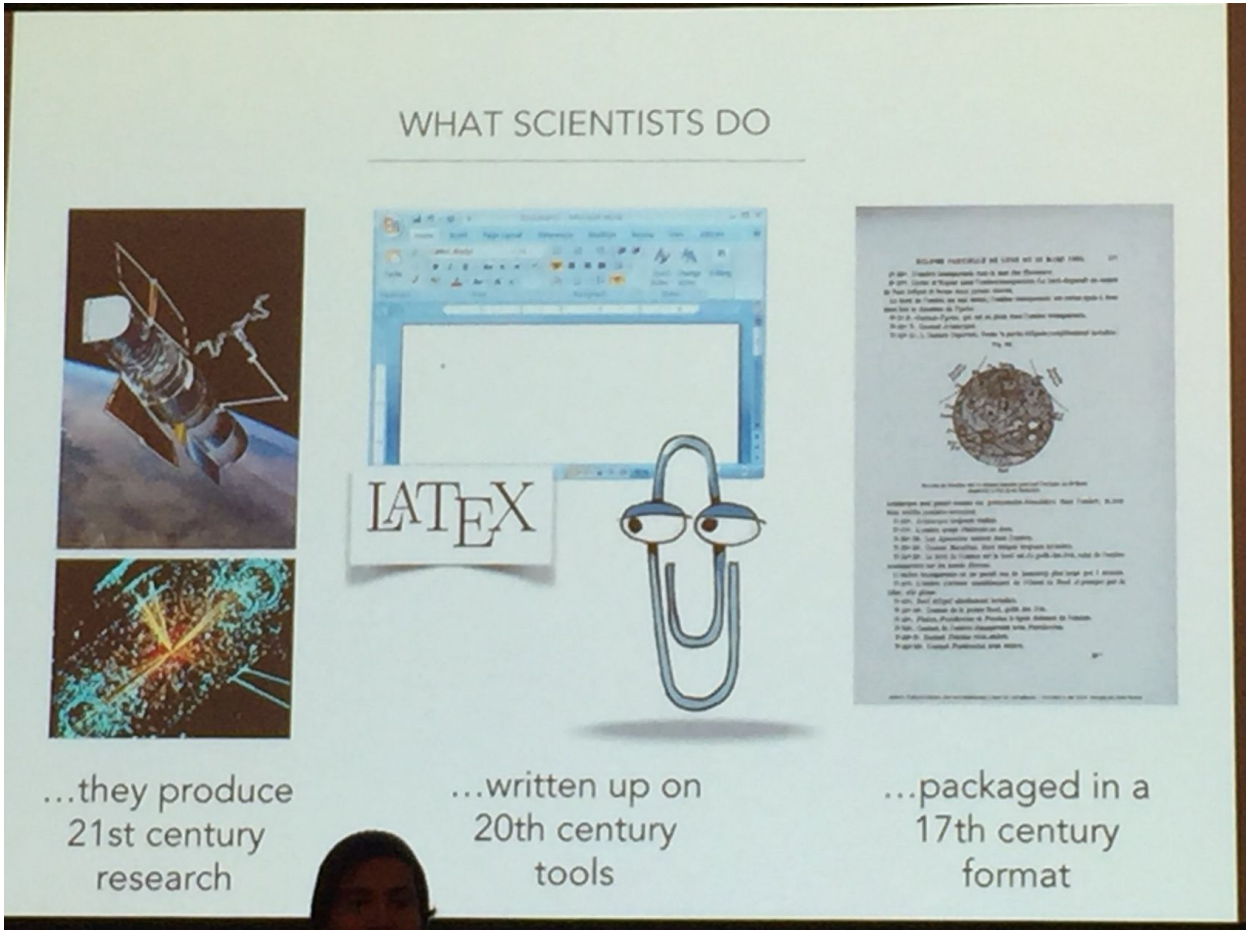
from

http://r4ds.had.co.nz/vectors.html#lists-of-condiments

Jenny Bryan (awesome explanations!)


1952
80
60
40
lifeExp
1e+03
1e+04
1e+05
gdpPercap
continent
Africa
Americas
Asia
Europe
Oceania
pop
2.50e+08
5.00e+08
7.50e+08
1.00e+09
1.25e+09

gganimate



Attempt to improve research tools + practices

Linear Modeling

All models are wrong, but some are useful.

- *George Box*

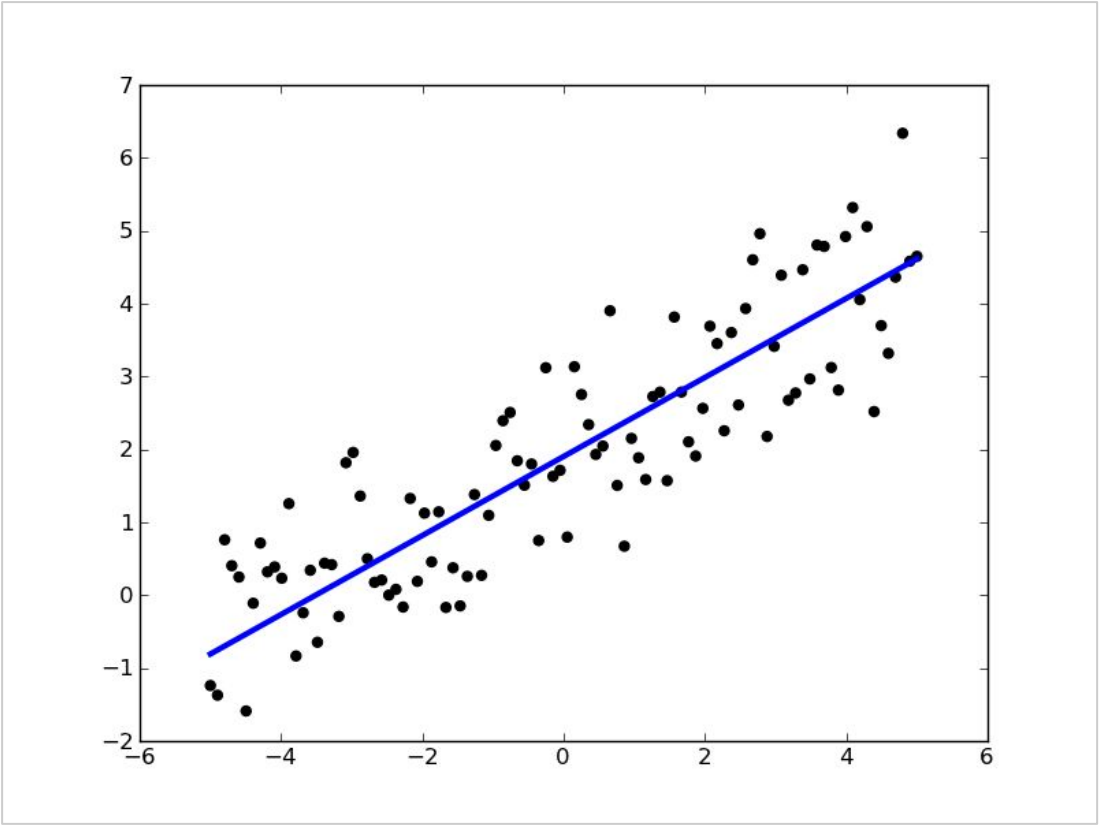
Linear Models

Assess the relationship between (1) dependent variable and (1+) independent variables

Allow you to describe association between features in your data

Assume a **linear** relationship (given a unit increase in X, expect a unit increase in Y)

There are entire courses (at UW!) on linear modeling



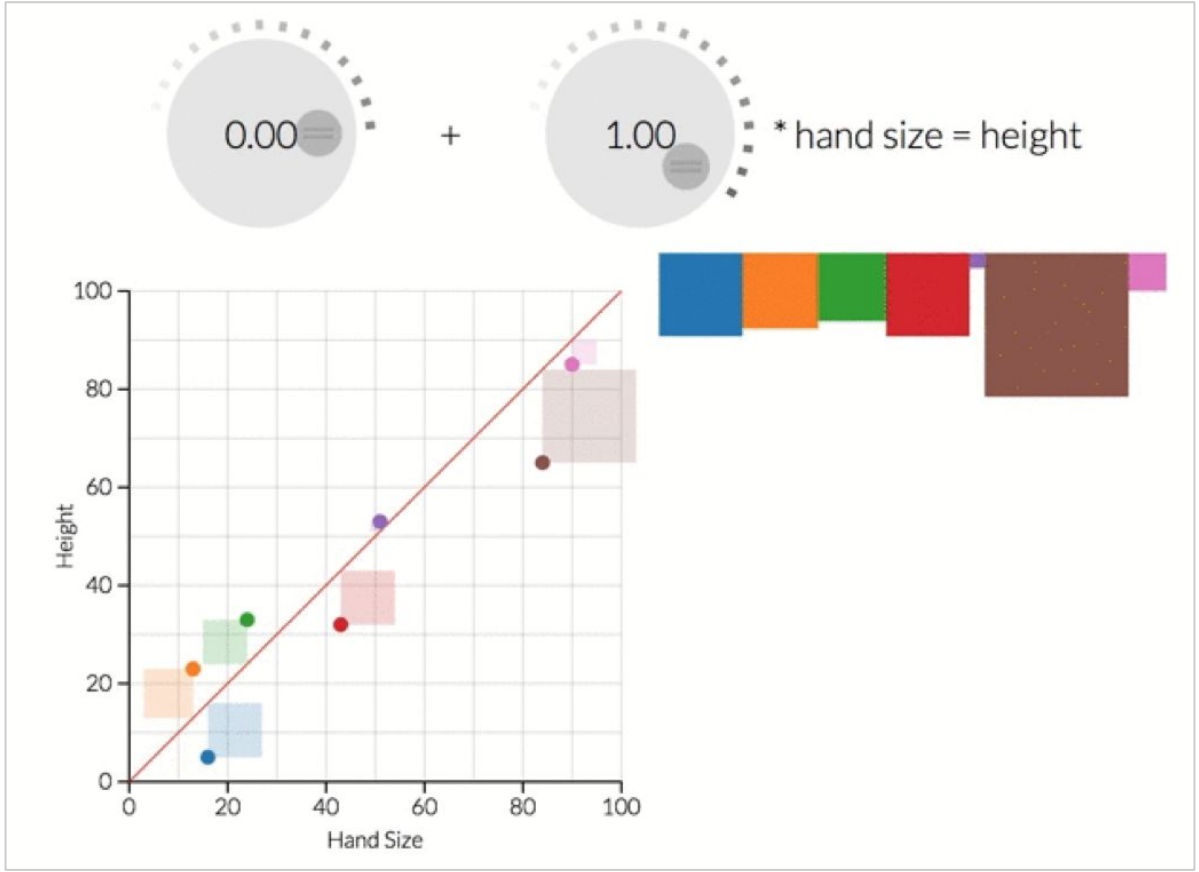
Find a function that describes the relationship between two variables by minimizing the squared distance between data and function

$Y = mx + b$

$$m = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Linear regression (ordinary least squares)

Minimizing squared distance???



Minimizing squared distance!

```r
# Generic
my.model <- lm(dep.var ~ ind.var, data=data)

# Example
library(gapminder)
lifeExp.model <- lm(lifeExp ~ year, data = gapminder)

# Better yet
gapminder$yearsElapsed <- gapminder$year - 1952
lifeExp.model <- lm(lifeExp ~ yearsElapsed, data = gapminder)
```

Managing Models

A terrible idea...?

Intentionally store dataframes in columns of dataframes

Helps keep data organized

Built on the idea that each row is a unit of analysis

Enables dataframe manipulation across rows (filter, select, etc.)

Might be a terrible idea

```r
install.packages('wesanderson')
install.packages('purrr')
install.packages('listviewer')
library(wesanderson)
library(purrr)
library(listviewer)

# Investigate with the structure (str) command
str(wesanderson)

# View with the jsonedit function (listviewer package)
jsonedit(wesanderson)

# Store names as a dataframe
wes.df <- data.frame(title=names(wesanderson))

# Add colors to dataframe
wes.df <- wes.df %>%
          mutate(colors = wesanderson,
                 num.colors = map(colors, length)
          )
```

Using the map functionality

module 15 exercise-1

Upcoming...

No lab this week

Have a great break!