Why R?

Reasons and applications for learning R

Matthias Raess

PhD candidate, applied linguistics

Department of English

Ball State University

mraess@bsu.edu / @primesty22

www.matthiasraess.com

What is R anyways??

- R is an implementation of the S programming language
- Developed by Ross Ihaka and Robert Gentleman at the University of Auckland, NZ
- Currently overseen by the *R Development Core Team*
- Project conceived in 1992, initial version 1995, stable beta in 2000.

Why R?

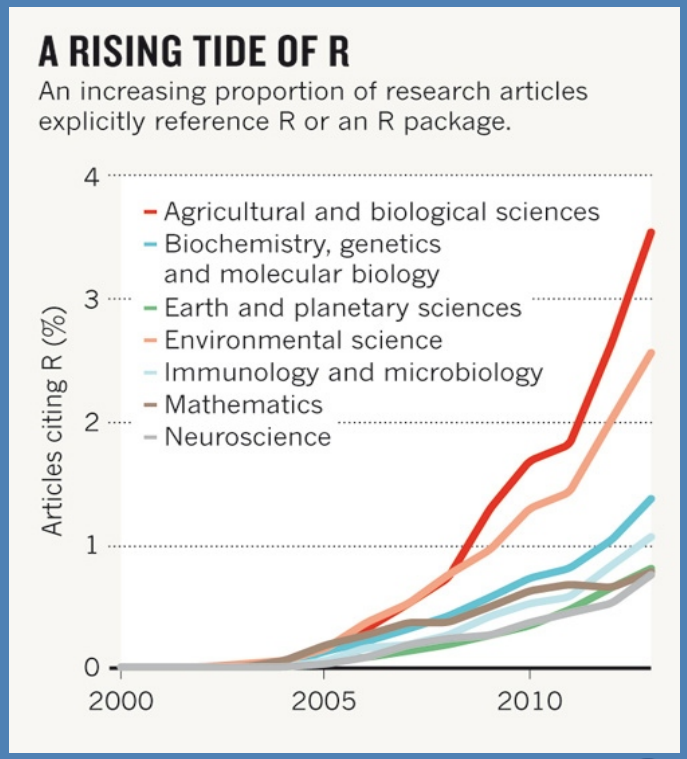
- Advantages: Platform independent (Windows/Mac/Linux)
- Open source > Changeability, big community > lots of materials
- Awesome plotting functions (base plotting, ggplot2, lattice)
- New stats method > chances are they come out in R first
- Workflow!! Unlike SPSS (GUI), step-by-step documentation (cf. Rmarkdown)
- Connection to open science movement + reproducible research
- Integrated version control via Git and GitHub
- FREE!!

Why R cont'

- Social sciences entered the age of data science!!
- R is becoming the lingua-franca of data science
- According to Redmonk, R ranks 13th of ALL programming languages, the highest of any statistical programming language
- Make yourself more marketable: big companies like Facebook and Google have their data scientists use R (also Bank of America, Ford, TechCrunch, Uber, and Trulia among others)

Why R cont' II

- R is very popular in academia
- Recent article in *Nature* (Tippmann, 2015) devoted to use of R in academia >
- The best and brightest people are trained in R at university > increases importance of R in industry



Why I learned R

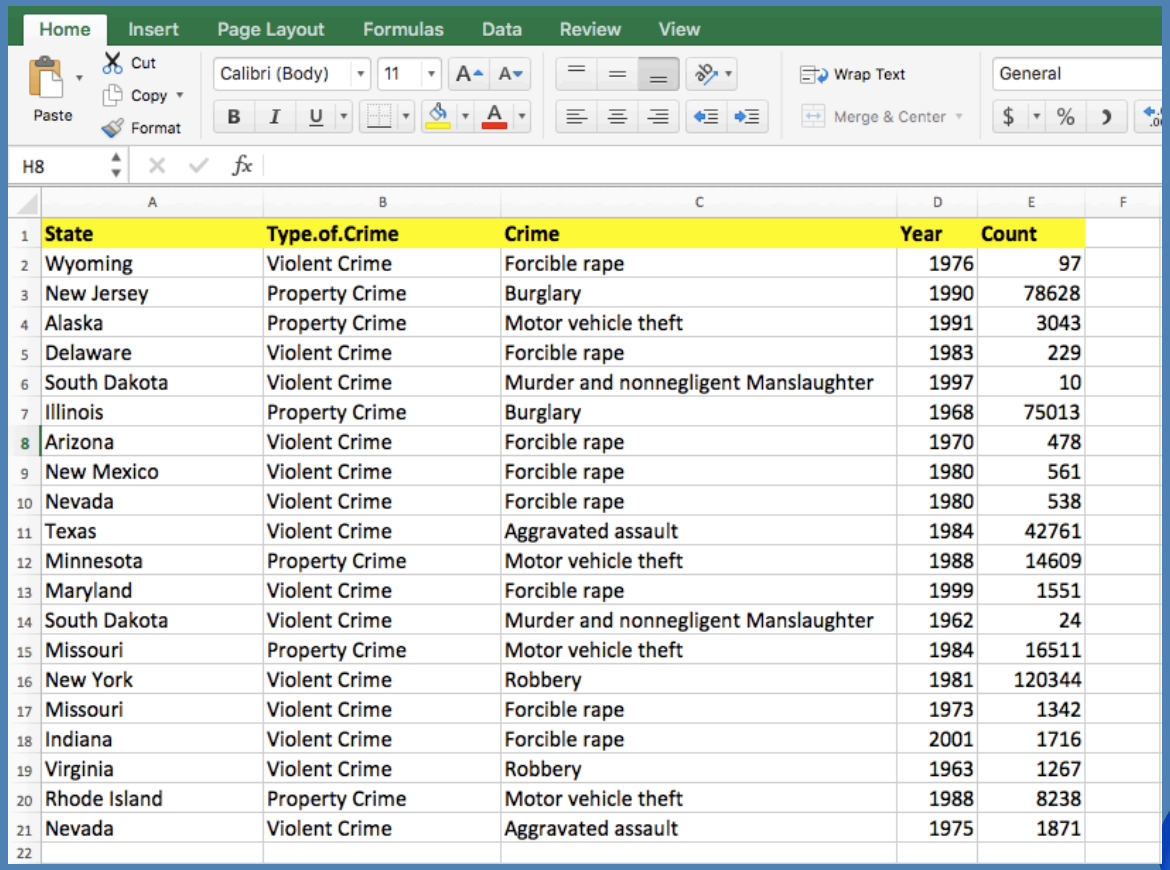
Learning R

- Lynda – Up and running with R – one of the best and ‘softest’ intros
- Lynda – R statistics essential training
- Code School – a very soft introduction to R
- DataCamp – Learn R for data science
- edX – Intro to R for data science – powered by Microsoft
- Coursera – Data science specialization (Johns Hopkins University) – the most time consuming option

Resources

- QuickR
- Rbloggers – R news and tutorials (580 contributors!!)
- STHDA – Just a ton of cool stuff about stats and R
- StackOverflow – This is where you ask a question if you have a problem!☺
- Rstudio - webinars by Rstudio
- Books: The R Cookbook (Teetor, 2011), Elegant Graphics for Data Analysis (Wickham, 2016), R for Excel Analysts (Taveras, 2016), R for Data Science (Wickham & Grolemund, 2017)

“Traditional” excel spread-sheet



State	Type.of.Crime	Crime	Year	Count
Wyoming	Violent Crime	Forcible rape	1976	97
New Jersey	Property Crime	Burglary	1990	78628
Alaska	Property Crime	Motor vehicle theft	1991	3043
Delaware	Violent Crime	Forcible rape	1983	229
South Dakota	Violent Crime	Murder and nonnegligent Manslaughter	1997	10
Illinois	Property Crime	Burglary	1968	75013
Arizona	Violent Crime	Forcible rape	1970	478
New Mexico	Violent Crime	Forcible rape	1980	561
Nevada	Violent Crime	Forcible rape	1980	538
Texas	Violent Crime	Aggravated assault	1984	42761
Minnesota	Property Crime	Motor vehicle theft	1988	14609
Maryland	Violent Crime	Forcible rape	1999	1551
South Dakota	Violent Crime	Murder and nonnegligent Manslaughter	1962	24
Missouri	Property Crime	Motor vehicle theft	1984	16511
New York	Violent Crime	Robbery	1981	120344
Missouri	Violent Crime	Forcible rape	1973	1342
Indiana	Violent Crime	Forcible rape	2001	1716
Virginia	Violent Crime	Robbery	1963	1267
Rhode Island	Property Crime	Motor vehicle theft	1988	8238
Nevada	Violent Crime	Aggravated assault	1975	1871

How R structures data

- The data frame (observations and variables)
- Tidy data paradigm (Wickham, 2014)
- > One row per observation, one column per variable (long format)
- Also
  - Vectors (a sequence of data elements of the same basic type)
  - and matrices (rows and columns)

Tidy data paradigm

Variable

Observation

State	Type.of.Crime	Crime	Year	Count
Wyoming	Violent Crime	Forcible rape	1976	97
New Jersey	Property Crime	Burglary	1990	78628
Alaska	Property Crime	Motor vehicle theft	1991	3043
Delaware	Violent Crime	Forcible rape	1983	229
South Dakota	Violent Crime	Murder and nonnegligent Manslaughter	1997	10
Illinois	Property Crime	Burglary	1968	75013
Arizona	Violent Crime	Forcible rape	1970	478
New Mexico	Violent Crime	Forcible rape	1980	561
Nevada	Violent Crime	Forcible rape	1980	538
Texas	Violent Crime	Aggravated assault	1984	42761
Minnesota	Property Crime	Motor vehicle theft	1988	14609
Maryland	Violent Crime	Forcible rape	1999	1551
South Dakota	Violent Crime	Murder and nonnegligent Manslaughter	1962	24
Missouri	Property Crime	Motor vehicle theft	1984	16511
New York	Violent Crime	Robbery	1981	120344
Missouri	Violent Crime	Forcible rape	1973	1342
Indiana	Violent Crime	Forcible rape	2001	1716
Virginia	Violent Crime	Robbery	1963	1267
Rhode Island	Property Crime	Motor vehicle theft	1988	8238
Nevada	Violent Crime	Aggravated assault	1975	1871

Types of variables

- Categorical variables
  - Also known as discrete or qualitative variables (nominal, dichotomous, ordinal)
  - E.g. gender, occupation, 0/1
- Continuous variables
  - Also known as quantitative variables (interval, ratio)
  - E.g. weight, height, age, income
- Some variables could be considered in either way.
  - E.g. Likert-scale items: attractiveness rating on 5-point scale continuous or categorical (5 levels)

Types of variables cont’

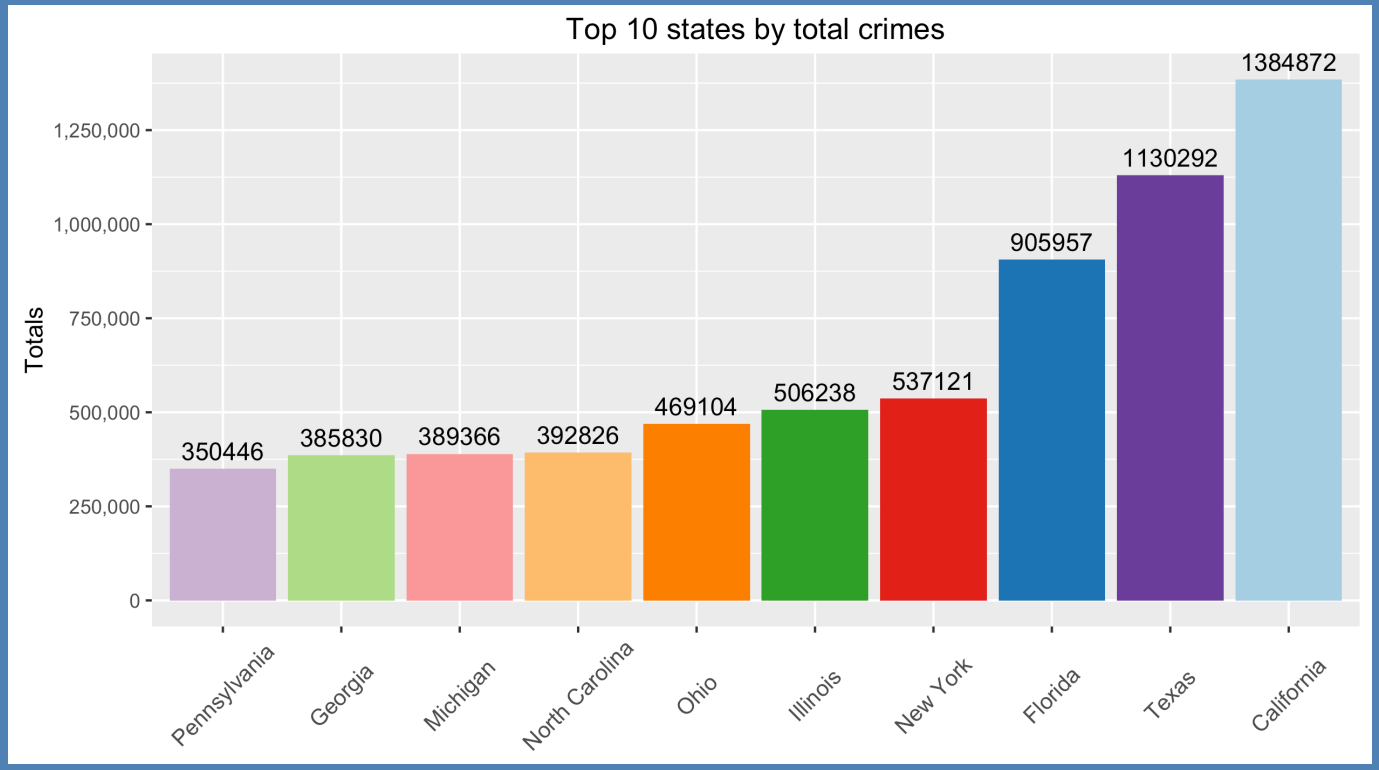
Categorical variable

Continuous variable

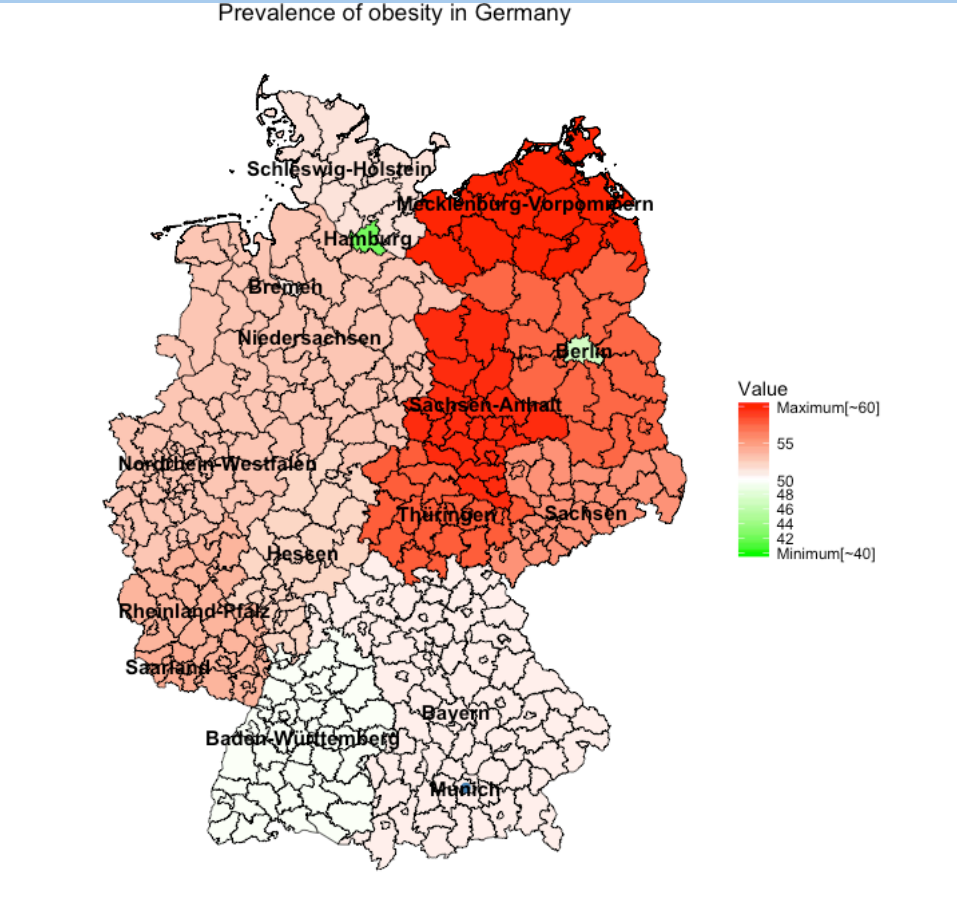
State	Type.of.Crime	Crime	Year	Count
Wyoming	Violent Crime	Forcible rape	1976	97
New Jersey	Property Crime	Burglary	1990	78628
Alaska	Property Crime	Motor vehicle theft	1991	3043
Delaware	Violent Crime	Forcible rape	1983	229
South Dakota	Violent Crime	Murder and nonnegligent Manslaughter	1997	10
Illinois	Property Crime	Burglary	1968	75013
Arizona	Violent Crime	Forcible rape	1970	478
New Mexico	Violent Crime	Forcible rape	1980	561
Nevada	Violent Crime	Forcible rape	1980	538
Texas	Violent Crime	Aggravated assault	1984	42761
Minnesota	Property Crime	Motor vehicle theft	1988	14609
Maryland	Violent Crime	Forcible rape	1999	1551
South Dakota	Violent Crime	Murder and nonnegligent Manslaughter	1962	24
Missouri	Property Crime	Motor vehicle theft	1984	16511
New York	Violent Crime	Robbery	1981	120344
Missouri	Violent Crime	Forcible rape	1973	1342
Indiana	Violent Crime	Forcible rape	2001	1716
Virginia	Violent Crime	Robbery	1963	1267
Rhode Island	Property Crime	Motor vehicle theft	1988	8238
Nevada	Violent Crime	Aggravated assault	1975	1871

ggplot2

- http://docs.ggplot2.org/current/index.html
- Written by Hadley Wickham – current version 2.2.0 (2016)
- Based on the grammar of graphics (Wilkinson, 2005) – plot made up of layers


Top 10 states by total crimes
Totals
1,250,000
1,000,000
750,000
500,000
250,000
0
350446
385830
389366
392826
469104
506238
537121
905957
1130292
1384872
Pennsylvania
Georgia
Michigan
North Carolina
Ohio
Illinois
New York
Florida
Texas
California

```
ggplot() +
    geom_polygon(data = merge.data,
aes(x = long, y = lat, group = group, fill =
year2013), color = "black", size = 0.25) +
coord_map() +
scale_fill_gradient2(name="Value", limits=c(40,60),
low="green", mid = "white", high="red", midpoint =
50, breaks=c(40,42,44,46,48,50, 55, 60),
labels=c("Minimum[~40]",42,44,46,48, 50, 55,
"Maximum[~60]"), space = "Lab") +
theme_nothing(legend = TRUE) +
labs(title = "Prevalence of obesity in Germany") +
geom_text(data = statenames, aes(long, lat, label =
states), size = 4, fontface = "bold",
col = "black", check_overlap = TRUE) +
geom_point(data = dot, aes(long, lat), color =
"steelblue", size = 2) +
geom_text(data = dot, aes(long, lat, label = city),
size = 4, fontface = "bold", color = "black")
```



~/Dropbox/BSU/WhyR - Trials-and-tribulations - RStudio

WhyR.R | Untitled1* | 01-Visualization.R | WhyR.Rmd | SideBar.R

```
crimeCountYear <- crimeData %>%
        group_by(Year) %>%
        summarize(CountYear = sum(Count))

ggplot(crimeCountYear, aes(Year, CountYear)) +
        geom_line() +
        scale_y_continuous(labels = comma)
```

Console | R Markdown

~/Dropbox/BSU/WhyR/

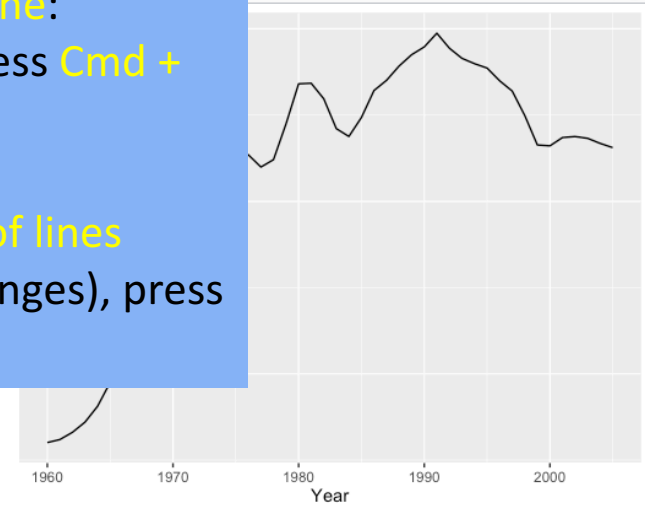
```
> head(crimeCountYear2)
Source: local data frame [6 x 3]
Groups: Year [3]

   Year  Type.of.Crime CountYear
  <int>         <fctr>     <int>
1  1960 Property Crime   2770689
2  1960  Violent Crime    238326
3  1961 Property Crime   2859039
4  1961  Violent Crime    238296
5  1962 Property Crime   3058865
6  1962  Violent Crime    247475
>
```

Environment | History | Git

Global Environment

Data

crime2000	10 obs. of 2 variables
crimeCountYear	46 obs. of 2 variables
crimeCountYear2	92 obs. of 3 variables
crimeData	16422 obs. of 5 variables



To "run" code in R, place the cursor on the line you want to run and press Cmd + Enter

Running more than one line: highlight the lines and press Cmd + Enter

To run the same "block" of lines again (e.g. with slight changes), press Cmd + Shift + P

~/Dropbox/BSU/WhyR - Trials-and-tribulations - RStudio

WhyR.R | Untitled1* | 01-Visualization.R | WhyR.Rmd | SideBar.R

```
crimeCountYear <- crimeData %>%
        group_by(Year) %>%
        summarize(CountYear = sum(Count))

ggplot(crimeCountYear, aes(Year, CountYear)) +
        geom_line() +
        scale_y_continuous(labels = comma
```

8:1 (Top Level)

Console | R Markdown

~/Dropbox/BSU/WhyR/

```
> head(crimeCountYear2)
Source: local data frame [6 x 3]
Groups: Year [3]

   Year  Type.of.Crime CountYear
  <int>         <fctr>     <int>
1  1960 Property Crime   2770689
2  1960  Violent Crime    238326
3  1961 Property Crime   2859039
4  1961  Violent Crime    238296
5  1962 Property Crime   3058865
6  1962  Violent Crime    247475
>
```

Environment | History | Git

Import Dataset | List

Global Environment

Data

crime2000	10 obs. of 2 variables
crimeCountYear	46 obs. of 2 variables

WhyR

Favorites: iCloud Drive, Desktop, Matthias, Dropbox, Documents, Applications, Public, Downloads, Box Sync, Google Drive

Devices: Matt SSD2, Matt SSD

Name	Date Modified	Size	Kind
01-Visualization.R	Oct 21, 2016, 7:17 AM	7 KB	Rez
AvgHP.png	Today, 9:07 PM	157 KB	PNG
crimePlot.png	Today, 9:07 PM	193 KB	PNG
crimeSpread.xlsx	Nov 4, 2016, 2:23 PM	9 KB	Micr
SideBar.R	Today, 4:41 PM	3 KB	Rez
spread2.xlsx	Nov 1, 2016, 12:41 PM	15 KB	Micr
testData.RData	Nov 1, 2016, 12:41 PM	562 bytes	R Da
testData.rds	Nov 1, 2016, 12:41 PM	538 bytes	R Da
untidyDat.xlsx	Nov 4, 2016, 2:38 PM	10 KB	Micr
WhyR.html	Today, 9:07 PM	2 MB	HTM
WhyR.R	Today, 9:11 PM	6 KB	Rez
WhyR.Rmd	Today, 9:07 PM	7 KB	R Ma
WhyR.Rproj	Oct 21, 2016, 11:49 AM	205 bytes	R Pro

Matt SSD › Users › Matthias › Dropbox › BSU › WhyR

13 items, 113.06 GB available

Publish