

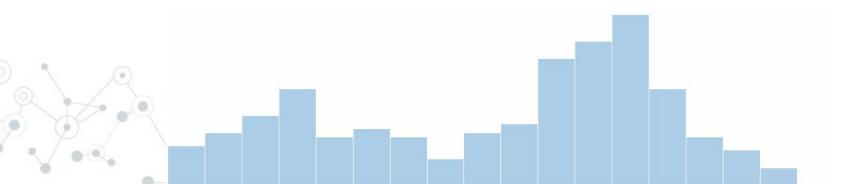
Introduction to





What is R?

- One of the most widely used data analysis software, used by statisticians, analysts, data scientists, etc.
- Powerful statistical programming language with unique data visualizations
- More than 14,000 libraries approved on CRAN (plus others on GitHub, etc.)
- R has more than 2 million users worldwide and is growing rapidly
- © R can be downloaded online for free along with **Rstudio**



How does R compare to other statistical software?

	SPSS	stata	S Sas	R	
Ease of learning	$\checkmark\checkmark$	\checkmark	\checkmark	×	
Good user interface	$\checkmark\checkmark$	✓	\checkmark	×	
Programming Capabilities	×	× v		$\checkmark\checkmark$	
Support from company	V	V	\checkmark	×	
Price	×	×	×	\checkmark	
Advanced Visualization capabilities	×	×	×	$\checkmark\checkmark$	
Handle complex models	×		$\checkmark\checkmark$	$\checkmark\checkmark$	
Handle large sets of data		\checkmark	$\checkmark\checkmark$	$\checkmark\checkmark$	

Rstudio

P

Be Edit Code yee Pipts Session Broket Build Sools Belo O - Off - The Edit of the Control of Control Build Sools Belo	📱 Project (bione) •
K. C. BER D. C. CHINARD	👰 uelig geod •
Paul R = _ R data sets = P Edd(25 Annua R = P Edd(21 - H Enes In, statter R =	
O I II F Source on Save Q. Z. O I I F Los and Running with R	
2 # £x05_02	Deta google 51 obs. of 9 variables
3 # Creating scatterplots 4 # how 2 quantitative variable associate with each other	
5 setwd("C:\\carmen\\R\\lynda")	territorial 1 (12) (3) Workspace
<pre>6 # Load data file about Google searches by state 7 google <- read.csv("google_correlate.csv", header = Y)</pre>	regl ln[11]
9 str(google)	
ripts there an association between the percentage of people	
17 + data of solization?	
17. # data visualization? 14. plotigoogleidegree, googleidata_viz)	
15 # Add title, labels, change circles to points .	_
<pre>16 plot(googleidegree, googleidata_viz, 17 main = "Interest in Data Visualization Searches\nby Percent of Population with C</pre>	
18 xlab = "Population with college begrees", 19 ylab = "Searches for \"Cata visualization"",	
20. pch = 20,	
21 col = "groy") 22 # Add fit lines	
23 # Linear regression line (y = x) 24 #	2
331 🔯 (Top Level) / R Script	The second se
Cansole Culument/Masslat/ 10	Files Plots Packages Help
\$ facebook : num 1.93 -0.52 -1.18 2.21 -1.28 -1.33 -0.14 -0.34 -2.2 0.1	🛃 🗢 🗇 🖉 Zoon 🖉 Equal - 🧕 🖉 Clear All
\$ nba : num -0.9 -1.08 1.23 -1.1 1.65 -0.68 0.1 0.03 -0.06 0.29 \$ has_nba : Factor w/ 2 levels "no","yes": 1 1 2 1 2 2 1 1 2 2	Interest in Data Visualization Searches
<pre>\$ degree : num 22.3 25.5 28 18.8 31.7 35.5 34.5 26.9 45.7 26 \$ stats_ed : Factor w/ 2 levels "no", "yes": 1 1 1 2 2 1 1 1 2 2</pre>	by Percent of Population with College Degrees
\$ region : Factor w/ 4 levels "Midwest", "Northeast",: 1 4 4 3 4 4 2 3 3 3	
# Is there an association between the percentage of people	
<pre>> # in a state with college degrees (%) and interest in > # data visualization?</pre>	Visualization
<pre>> plot(googleidegree, googleidata_viz) > # Add title, labels, change circles to points</pre>	
<pre>> plot(googleSdegree, googleSdata_viz,</pre>	07 N -
 main = "Interest in Data visualization Searches\nby Percent of Population with College Degrees". 	5
 Xiab = "Population with college begrees", ylab = "Searches for \"Data visualization\"". 	Data
oncolo	2
,onsole	
> # Linear regression 1 re (y = x)	
> abline(lm(googleidata_viz = googleidegree), col="red")	4) Results/F
<pre>> # Lowess smoother line (x, y) > # might now be linear, so want to draw a line that's</pre>	
> # closer to the shape of the data > # note = order is different	
> # INTE = CLOBE IS CULLERAIL	15 20 25 30 35 40 45
> # Towess = moving average	



Is R right for you?

Advantages

- Open-Source
- Community support
- Automation
- Flexibility
- Dynamic output

Disadvantages

- Steep learning curve
- Programming and capacity limitations when compared to Python or similar
- Some libraries may not be updated
- Not standardized

Using R to Work with Census Data

R allows you to download census data directly. <u>Steps:</u>

- 1. Request a free Census Bureau API key <u>https://api.census.gov/data/key_signup.html</u>
- Download a few packages: tigris (shapefiles), tidycensus (Census and ACS data with feature geometries) and sf, (simple features is use to represent geographic vector data).



3. Load variables of interest.

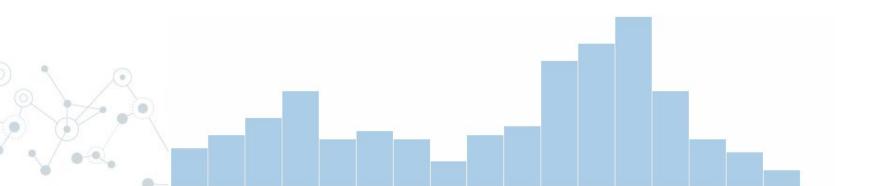
```
nc_pop <-
get_acs(geography = "county",
            variables = "B01003_001",
            state = "NC",
            geometry = TRUE)</pre>
```

4. Your are now ready to interact with the data

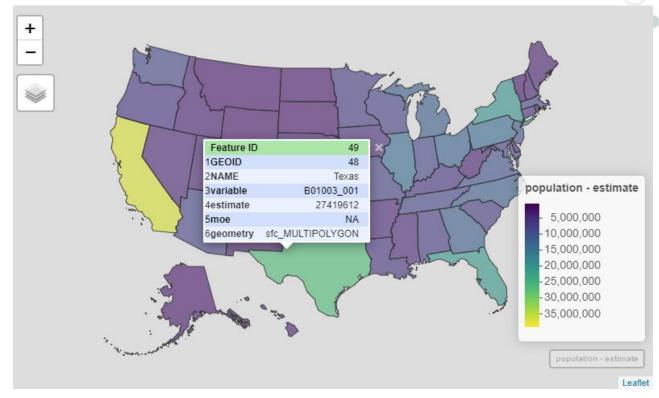
Continuation of Census Data and R

If we install the **leafview** and **mapview** packages, we can visualize the data:

```
mapviewOptions(legend.pos = "bottomright")
mapviewOptions(leafletWidth = 800)
#mapviewOptions()
#mapviewOptions(default = TRUE)
mapview(population, zcol = "estimate", native.crs = TRUE, crs = 5070)
```



Example Output



https://map-rfun.library.duke.edu/02_choropleth.html

Using R for Survey Data - Jamie

Using libraries **gmodels** and **wordcloud**, R can analyze frequencies, cross-tabs and text.

> CrossTable(survey\$Q7, survey\$Q2, prop.t=FALSE, prop.r=TRUE, prop.c=FALSE, prop.chisq=FALSE, format="SPSS", digits=1)

 	Count Row Percent					
otal Observat	ions in Table	45				Se testing
 survey\$Q7	survey\$Q2 A	в	c	D	Row Total	education position programs
A	10.0%	7 70.0%	2 20.0%	0 0.0%	10 22.2%	
В	12 37.5%	14 43.8%	6 18.8%	0 0.0%	32 71.1%	
C	2 66.7%	0 0.0%	1 33.3%	0 0.0%	3 6.7%	special ESE CMS
column Total	15	21	9	0	45	

systems also cumbersome eclalists given by the cumbersome support sees an also cumbersome difficult there consum needed things

Data Manipulation



Derive new variables



Join multiple data sets of data together



Create summaries of your dataset



Pull information directly from websites and/or public data sets (e.g. ACS)

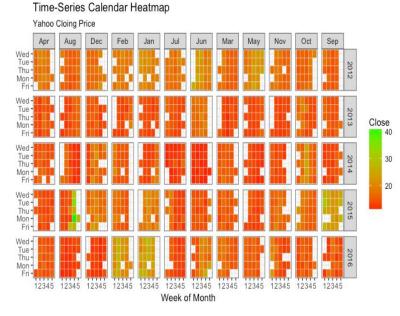




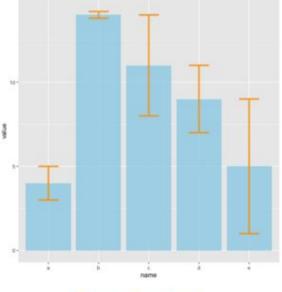
Data Visualization

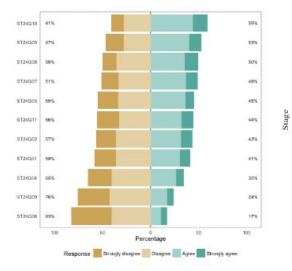
R has several packages that enable visualizing data:

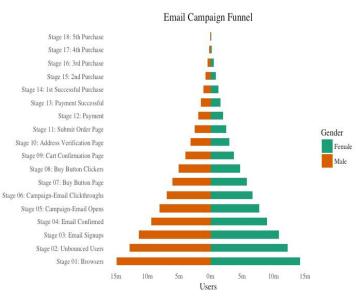
- BaseR
- Ggplot2
- Leaflet (interactive)
- Plotly (interactive)
- Other specialized (various models, EDA, GIS, network, etc.)





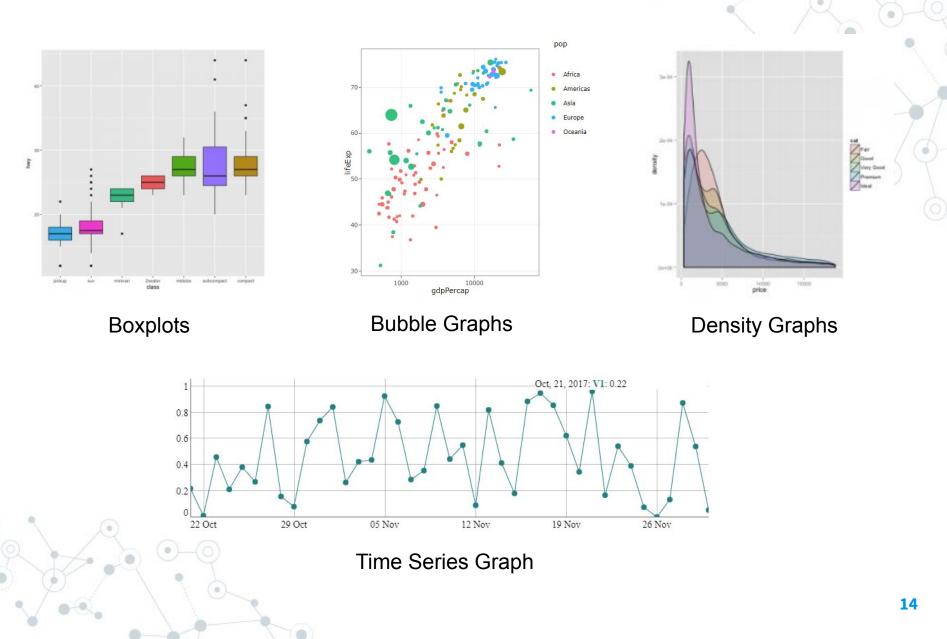




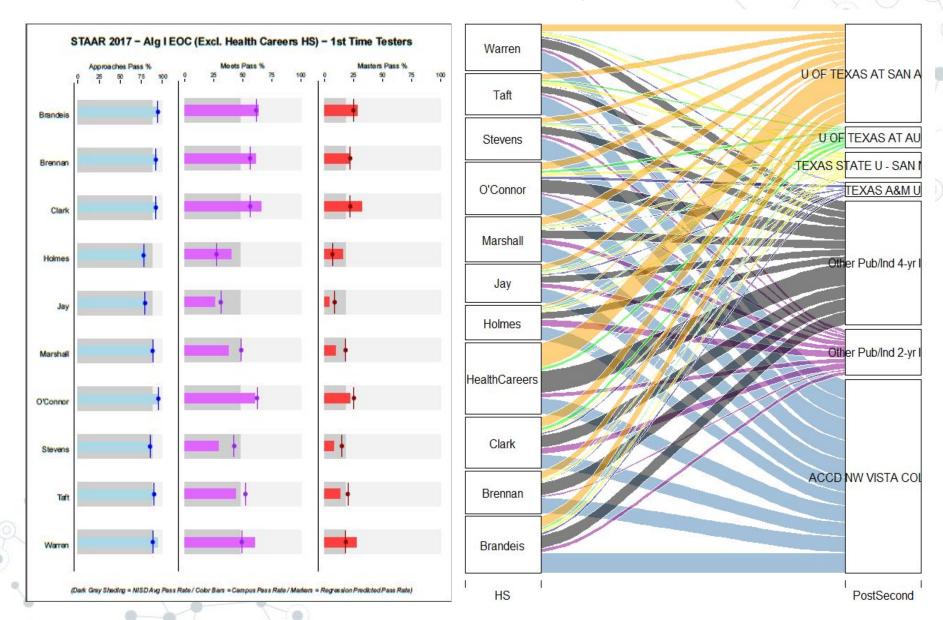


#4 Error bars on barplot

Data Visualization



EXAMPLES OF PROJECTS Visualizations of STAAR Results & College Enrollment Flows



EXAMPLES OF PROJECTS Decision Trees using CTREE

SAT SY1314 READING

race_eth2 p < 0.001 {W, A} (B, H). -2 eco_dis p < 0.001 eco_dis p < 0.001 -6 race_eth2 p = 0.024Node 4 (n = 188) Node 7 (n = 1248) Node 3 (n = 983) Node 8 (n = 182) Node 9 (n = 1226) 800 800 800 800 800 600 600 600 600 600 400 400 400 400 400 8 200 200 -200 -200 200 0 0 0 0 0

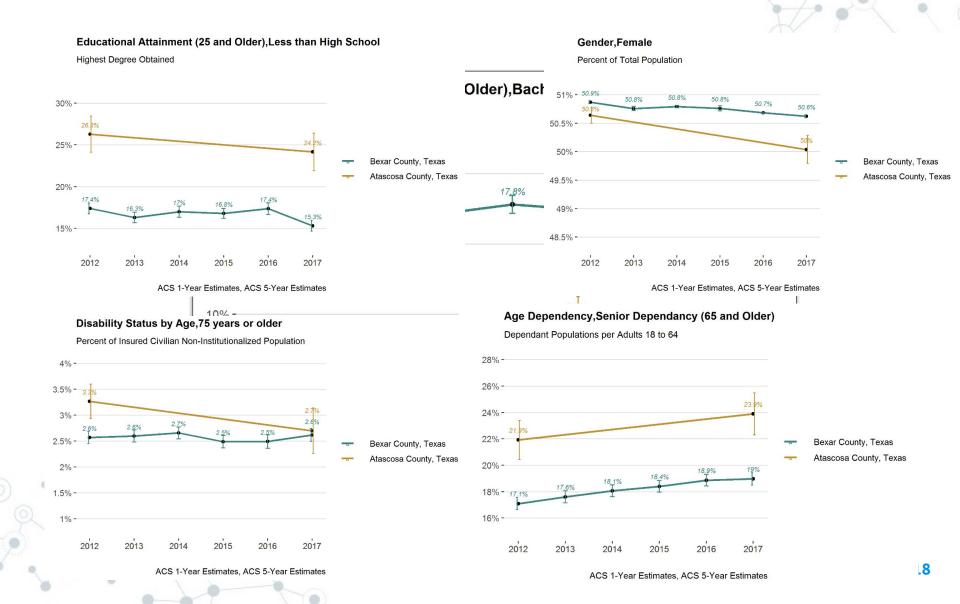
SAT SPR 2015 READING EcoDis p < 0.001 Race_Eth2 -12 Race_Eth2 p < 0.001 p < 0.001 {W, A} {A, B, H} {B, H} Node 3 (n = 1243) Node 6 (n = 290) Node 7 (n = 2316) Node 4 (n = 2355) 800 800 800 -800 600 600 600 600 400 400 400 400 200 200 200 200 0 0 0 0 0

EXAMPLES OF PROJECTS

Automation and customization of over 200 Trendlines

						10 - C - C - C - C - C - C - C - C - C -		
Geoid	Title	Subtitle	Source	Year	Estimate	Margin of Error (Moe)	Min Moe	Max Moe
Atascosa County	Educational Attainment (25 and Older), Bachelors Degree	Highest Degree Obtained	ACS 1-Year Estimates, ACS 5-Year Estimates	2012	8.15	1.38	6.77	9.53
Atascosa County	Educational Attainment (25 and Older), Bachelors Degree	Highest Degree Obtained	ACS 1-Year Estimates, ACS 5-Year Estimates	2017	9.61	1.64	7.97	11.25
Bexar County	Educational Attainment (25 and Older), Bachelors Degree	Highest Degree Obtained	ACS 1-Year Estimates, ACS 5-Year Estimates	2012	16.5	0.59	15.91	17.10
Bexar County	Educational Attainment (25 and Older), Bachelors Degree	Highest Degree Obtained	ACS 1-Year Estimates, ACS 5-Year Estimates	2013	17	0.58	16.42	17.58

EXAMPLE OF PROJECT: AUTOMATION AND CUSTOMIZATION OF TRENDLINES



```
```{r}
1 -
 trendpercent$geoid <- factor(trendpercent$GEOID, levels = c("Bexar County, Texas","Atascosa County, Texas"))</pre>
 2
 # create graphing function
 4 - trend.graph <- function(trendpercent, na.rm = TRUE, ...){
 5
 6
 # create list of title in data to loop over
 Create an Index
 7
 title_list <- unique(trendpercent$CHARTTITLE)</pre>
 8
 9
 # create for loop to produce ggplot2 graphs
10 -
 for (i in seq_along(title_list)) {
11
 Y Axis
12
 # create plot for each title in trend and adds values for x and y plotting
13
 plot <-
 X Axis
14
 gqplot(subset(trendpercent, trendpercent$CHARTTITLE ==title_list[i]),
15
 aes(x=YEAR, y=ESTIMATE, ymin=minmoe, ymax=maxmoe, group=(GEOID), color=GEOID)) + theme_hc()
 Group
16
 Min Moe
 geom_line(aes(colour=GEOID),size = 1.0) + geom_point(color="black")+
17
18
 Max Moe
 #labels in bars, paste % symbol, adjust color and position
19
20
 geom_text_repel(aes(label = paste0(round(ESTIMATE,1),"%")), size=2.5 , fontface='italic', force=1,
 Labels
21
 segment.color = 'transparent', min.segment.length = 0, segment.size = 0, point.padding = .3,
22
 nudge_x = 0, nudge_y = .1, direction = "y")+
23
 #add the % symbol to the values in axis and adjust limits to be automated and adds the function "expand".
24
25
 scale_y_continuous(labels=function(ESTIMATE) paste0(ESTIMATE,"%"), expand = expand_scale(mult = c(.1,.3),
 Limits
26
 0)),breaks=pretty_breaks(n=5),limits=c(min(trendpercent$minmoe[trendpercent$CHARTTITLE ==title_list[i]])
 add = c(1.5.
27
 ,max(trendpercent$maxmoe[trendpercent$CHARTTITLE ==title_list[i])))+
 and
28
 breaks
29
 scale_x_continuous(breaks=c(trendpercent$YEAR))+
30
31
 #specify colors for lines
 Color of
32
 scale_color_manual(values=c("Bexar County, Texas"="#3c857f", "Atascosa County, Texas"="#c09231"))+
33
 lines
34
 #modify leaend
35
 theme(legend.position="right", legend.title = element_blank(),legend.spacing.x = unit(.6, 'cm'),
 Legend
36
37
 #remove panel, set background color and reorder items in legend
38 Background
 panel.grid.major = element_blank(), panel.background = element_rect(fill = "white")) +
39
40
 #error bars
41
 Error bars
 geom_errorbar(aes(ymin=minmoe, ymax=maxmoe), width=.1, position = position_dodge(0.1)) +
42
43
 #automates title and specified x and y labels as blank
 ggtitle(title_list[i])+labs(x="", y="", caption = trendpercent$SOURCEID [trendpercent$CHARTTITLE ==title_list[i]],
44
 Title,
45
 subtitle = trendpercent$CHARTSUBTITLE[trendpercent$CHARTTITLE ==title_list[i]]) +
 theme(plot.title=element_text (hiust = 0, face='bold', size=12, margin=margin(0,0,10,0))) +
46
 subtitle.
 theme(plot.subtitle = element_text(hjust=0, size=10,margin=margin(0,0,20,0)))+
47
48
 theme(plot.margin = unit(c(.5,.9,.5,.5), "cm"))#top, right, bottom, left
 source
49
50
 Save as an
 #Export as images
51
 gqsave(paste0(title_list[i],".png"))
 image
52
53
 # print plots to screen
54
 print(plot)
55
 331
56
 # run graphing function
57
 trend.graph(trendpercent, ESTIMATE)
E O
```

### Resources



#### **News & Tutorials**

#### **R-bloggers**

Blogs related to R and its applications

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

#### **R Graph Gallery**

Examples of visualizations with code samples

https://www.r-graph-gallery.com/

### Books



#### Rdocumentation

Manuals and information for packages https://www.rdocumentation.org/

#### Stackoverflow

Developers share knowledge https://stackoverflow.com/



#### Rtips

List of common tasks performed in R http://pj.freefaculty.org/R/Rtips.html.

#### ImpatientR

Introduction to basic functions

https://www.burns-stat.com/documents/ tutorials/why-use-the-r-language/

#### YaRrr! The Pirates Guide to R

Intro to basic analytical tools in R, from basic coding and analyses, to data wrangling, plotting, and statistical inference.

https://bookdown.org/ndphillips/YaRrr/





# **Thanks!** Questions?

### **Paulina Cano:**

paulina.canomccutcheon@uth.tmc.edu

#### Jamie Ford:

jamie.ford@nisd.net



