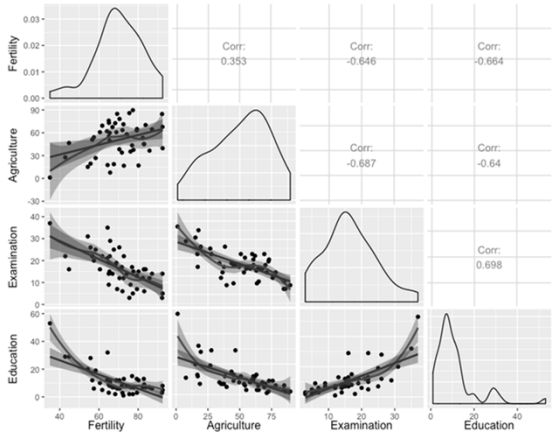
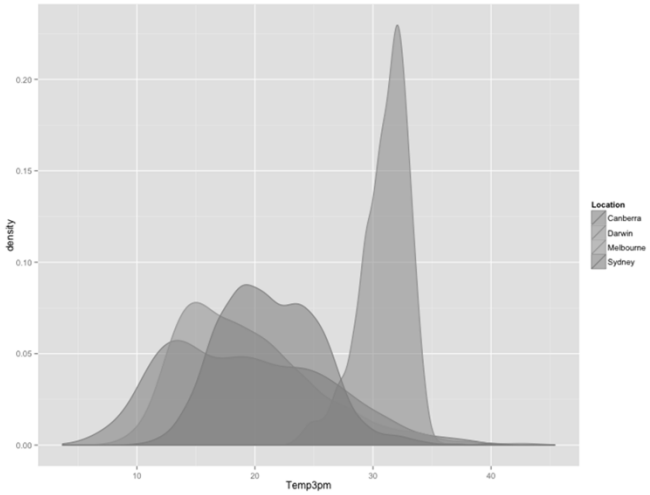
#DataVisualization in R



Brady T. West, Ph.D.
Research Associate Professor
Survey Research Center
Institute for Social Research
University of Michigan-Ann Arbor
bwest@umich.edu



Webinar Overview

- Five Key Aspects of Effective #DataViz:
  - **Who** are we talking about? (paint a picture!)
  - **What** are we trying to visualize? (what is the statistic?)
  - **When** were the data collected? (incorporate time!)
  - **Where** were the data collected? (incorporate geography!)
  - **Why** were there certain outcomes? (visualize associations!)
- Examples of each approach using the R / RStudio software
- Additional references / textbooks / articles / web pages
- Go here for the slides (and hence the R code): http://www.umich.edu/~bwest/dataviz-webinar.pptx

Some survey data from NHANES…

- We'll start by importing some survey data into R
- These data are from the 2011-2012 National Health and Nutrition Examination Survey (NHANES)
- **Why NHANES?** While clearly not all surveys arise from probability samples, effective visualization of population features needs to consider the role that **weights** may play in generating graphics
- There are also many types of variables available in NHANES
- We will consider procedures appropriate for either weighted or unweighted survey data

Importing the data (RStudio)

```
load(url("http://www-personal.umich.edu/~bwest/nhanes1112_webinar.rdata"))

dim(nhanes1112_sub_10jun2016) # to get a sense of cases and variables
[1] 9756    39

summary(nhanes1112_sub_10jun2016)
# a variety of continuous and categorical variables, including weights (WTMEC2YR):

      seqn          ridstatr        riagendr        RIDRETH1        dmdmartl
 Min.   :62161   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   : 1.000
 1st Qu.:64600   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:3.000   1st Qu.: 1.000
 Median :67039   Median :2.000   Median :2.000   Median :3.000   Median : 2.000
 Mean   :67039   Mean   :1.957   Mean   :1.502   Mean   :3.229   Mean   : 2.749
 3rd Qu.:69477   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:4.000   3rd Qu.: 5.000
 Max.   :71916   Max.   :2.000   Max.   :2.000   Max.   :5.000   Max.   :99.000
                                                                 NA's   :4196
    WTINT2YR         WTMEC2YR         sdmvpsu         sdmvstra        indfmpir
 Min.   :  3321   Min.   :     0   Min.   :1.000   Min.   : 90.00   Min.   :0.000
 1st Qu.: 11352   1st Qu.: 11174   1st Qu.:1.000   1st Qu.: 92.00   1st Qu.:0.860
 Median : 18098   Median : 18090   Median :2.000   Median : 96.00   Median :1.630
 Mean   : 31426   Mean   : 31426   Mean   :1.643   Mean   : 95.87   Mean   :2.205  ...
 3rd Qu.: 34887   3rd Qu.: 34792   3rd Qu.:2.000   3rd Qu.: 99.00   3rd Qu.:3.580
 Max.   :220233   Max.   :222580   Max.   :3.000   Max.   :103.00   Max.   :5.000
                                                                    NA's   :840
```

Declaring sample design features

- First, install and load the survey package:

```
install.packages("survey")
library(survey)
```

- Next, create a survey design object describing the sampling features:

```
nhanes1112_sub_10jun2016$riagendr <- factor(nhanes1112_sub_10jun2016$riagendr,
levels=c(1,2), labels=c("Male","Female")) # add value labels to gender

nhanes1112_sub_10jun2016$RIDRETH1 <- factor(nhanes1112_sub_10jun2016$RIDRETH1,
levels=c(1,2,3,4,5), labels=c("Mexican","Other Hispanic","White","Black","Other"))
# add value labels for race / ethnicity

nhanes.dsgn <- svydesign(id=~sdmvpsu, strata=~sdmvstra, weights=~WTMEC2YR,
data=nhanes1112_sub_10jun2016, nest=TRUE)

summary(nhanes.dsgn)
```

- Now, we can employ selected procedures that allow us to generate weighted graphics, describing target populations!
- Our main focus will be on graphics for unweighted data, but **you should always decide if weights are necessary for your visualization**

Principles of Effective Data Visualization

- **Let's start with some general principles of effective data visualization!**
- Parsimony in visualization should be the goal, but not to the point of eliminating important information
- **Is visualization even needed?** If you just want to talk about, for example, a simple difference in means between two groups, a graphic might not be necessary…what about differences in *variance* as well?
- Avoid unnecessary use of different colors (e.g., the title slide)
- Avoidclutteringinlabelsofdatapoints
- Clearly label all axes of a graph

Principles of Effective Data Visualization

- Assume that your audience is pretty bright! You can present a complex visualization, as long as everything is clear
- **Clarity is the absolute key**; if a scatter plot with different subgroups labeled with different colors tells the whole story, then so be it! Don't make graphs unnecessarily complicated.
- Avoid clutter / unnecessary formatting / overlap of text
- **Smart** use of color should not be optional; most people will read your work online, so effectively incorporate color into your graphics!
- Make sure that all text is clear and easy to read
- Make effective use of **transparency,** especially for overlaid elements

Some Baaaaaad #DataViz…What's Wrong?

h/t to my esteemed colleague @DetroitTexan

Principles of Effective Data Visualization

- If possible, replace histograms with density plots to illustrate distributions
- Avoid pie charts if possible (will a table suffice?)
- Avoid the unnecessary use of 3D graphics
- When plotting maps, show geographic variation in key variables, not just population densities!
- If you must add error bars to a plot, make sure to clearly indicate what they mean (SE? 2 x SE? CI? Something else?)

Principles of Effective Data Visualization

- Provide visual indications of variability (for a variable, or for an estimate) whenever possible / meaningful
- Provided that the result is clear and concise, don't hesitate to combine multiple charts into a single figure (easy to do in R; see https://t.co/gpTEhFhfY5 for more details!)
- When generating visuals, let colleagues proofread them, just like you would let them proofread reports and journal articles!
- **Now, let's turn to examples of visualizing who, what, when, where, and why!**

Who are we talking about?

- We'd like to paint a visual picture of our target population, represented by the survey data
- Simple displays of the distributions on socio-demographic variables are often helpful…
- …but don't be afraid to just report a clear table in simple cases
- Make sure to also indicate the sampling variability in all estimates
- Suppose that we wish to visualize the estimated distribution of race / ethnicity in the target NHANES population

Who are we talking about?

- First, generate the estimated distribution (and standard errors):

```
> svymean(~RIDRETH1, design = nhanes.dsgn, se = T, na.rm = T)
                            mean     SE
RIDRETH1Mexican         0.097238 0.0208
RIDRETH1Other Hispanic  0.069853 0.0154
RIDRETH1White           0.628595 0.0407
RIDRETH1Black           0.124373 0.0239
RIDRETH1Other           0.079942 0.0104
```

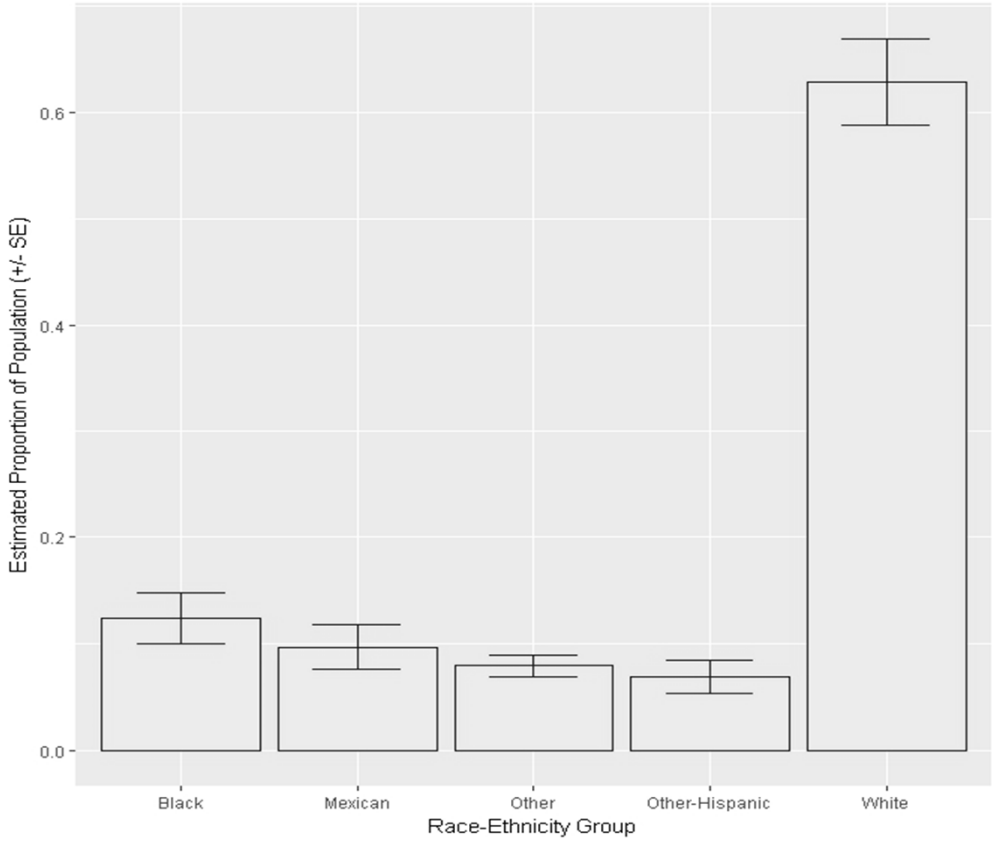
Who are we talking about?

- Now, create a data frame of estimates, load the ggplot2 package, and generate the bar chart with SE bars:

```
my.data <- read.table(text = "group est se
Mexican 0.0972 0.0208
Other-Hispanic 0.0699 0.0154
White 0.6286 0.0407
Black 0.1244 0.0239
Other 0.0799 0.0104", h=T)

library(ggplot2)
ggplot(my.data, aes(x = factor(group), y = est)) +
  geom_bar(stat = "identity", color = "blue", fill = "yellow") +
  geom_errorbar(aes(ymin=est-se, ymax=est+se), color = "blue", width = 0.5) +
  xlab("Race-Ethnicity Group") + ylab("Estimated Proportion of Population (+/- SE)")
```

Who are we talking about?



Who are we talking about?

- Consider visualizing a population in terms of the proportion of people with some characteristic, measured using a survey
- **Risk characterization theaters (RCTs)!**
- If a population has X cases that meet a certain criterion out of 1,000, consider using the RCT.R function (and corresponding .Rdata file) provided at these web sites:

https://github.com/seancarmody/stubborn-mule/tree/master/RCT/

https://www.r-bloggers.com/generate-your-own-risk-characterization-theatre/

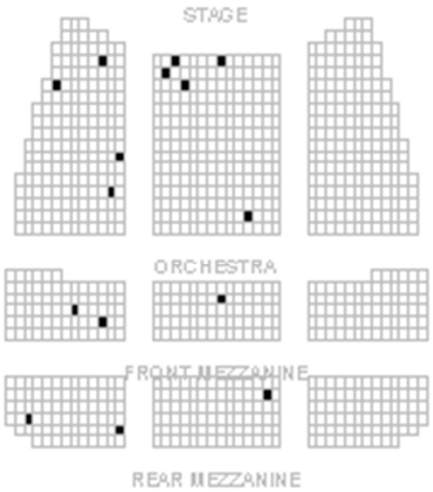
Who are we talking about?

- **Example:** A weighted analysis of the NHANES data suggests that 1.52% of males and 1.17% of females have an irregular heartbeat:

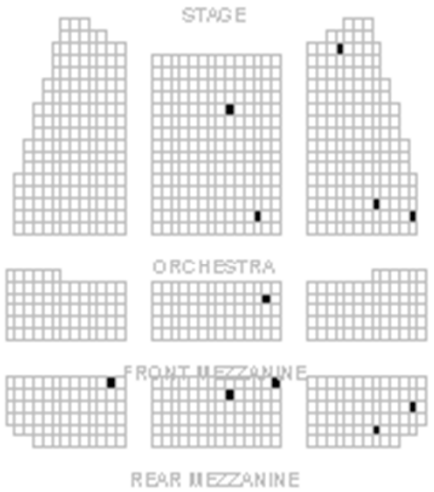
```
svyby(~irregular, ~riagendr, svymean, design = nhanes.dsgn, na.rm = T)
       riagendr  irregular          se
Male       Male 0.01521464 0.002635648
Female   Female 0.01166096 0.001590939

source(url("http://www-personal.umich.edu/~bwest/RCT.R")) # load RCT.R function
par(mfrow = c(1,2)) # note use of a 1 x 2 matrix of plots!
# use rct() to visualize the distribution (per 1,000) for each sex, theatre view
rct(15.2, type = "theatre", xlab = "Males", label = TRUE, lab.cex = 0.5)
rct(11.7, type = "theatre", xlab = "Females", label = TRUE, lab.cex = 0.5)
```

RCT with a theatre view (stadium also possible)!



Males



Females

What are we trying to visualize?

- It is essential to paint a clear picture of distributions on survey variables that we are trying to analyze!
- An incredibly hot and controversial topic is what has been happening with lead poisoning of the water supply in Flint, Michigan
- I consulted on a study published in *Pediatrics* to look at historical trends in the distribution of blood lead levels (BLLs) among children measured by the Hurley Medical Center in Flint
- We can also use the `ggplot2` package to create visually appealing comparisons of distributions across time / across groups

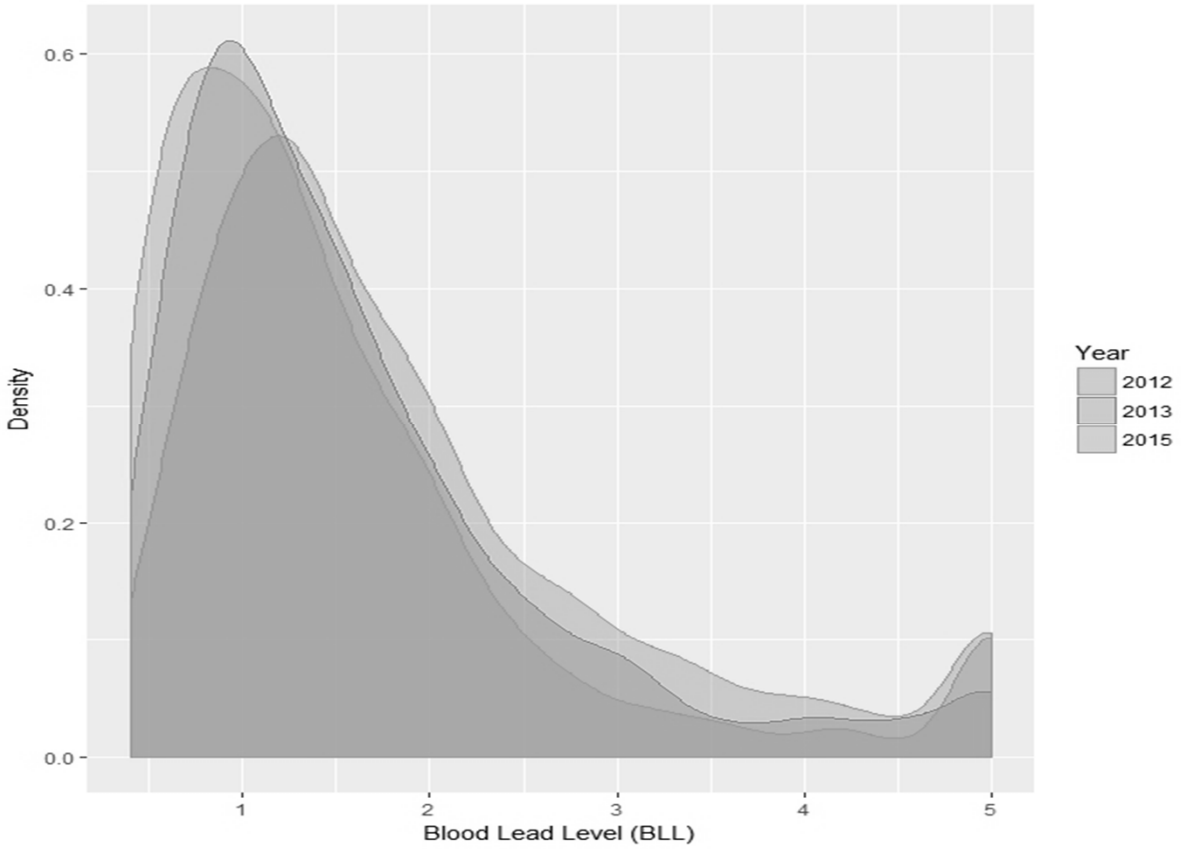
What happened in Flint with BLLs?

```
# load data (**not publicly available)
load("H:\\final_flint_data.rdata")
# load ggplot2 package
library(ggplot2)

# focus on data inside city limits
flint <- subset(final_flint_data, inside == 1)
flint$Year <- factor(flint$Year)
# top-code BLLs, per CDC
flint$bll[flint$bll >= 5] <- 5

# use ggplot to create overlaid density functions, with clear labels
ggplot(flint, aes(x=bll, color=Year, group=Year, fill=Year)) +
geom_density(alpha=0.4) +
xlab("Blood Lead Level (BLL)") + ylab("Density")
```

What happened in Flint with BLLs?



Attractive features of this visualization…

- The `ggplot2` package (gg = "grammar of graphics") provides many essential tools for visualization in R (but this is not the only package!)
- Note the use of **transparency**
- Note the effective use of **colors**
- Note the use of **density functions**
- Note the **clear labeling**
- A clear picture emerges in terms of differences in distributions, and specific points where the distributions vary; **and we don't need that much code to generate the visualization!**

What are we looking at?

- The `plotly` package is also an excellent tool for **interactive visualization** of distributions (RStudio recommended): see https://plot.ly/r/getting-started/#installation
- One popular choice is a **dot plot**; we consider the utility of such a plot for **visualizing an interaction** between race / ethnicity and gender in predicting the probability of an irregular heart beat (a binary variable)
- First, we analyze the data, create a data frame containing the estimates, and load the `plotly` package:

```
tempirr <- as.data.frame(svyby(~irregular, ~riagendr + RIDRETH1, svymean,
design = nhanes.dsgn, na.rm = T))

install.packages("plotly")

library(plotly)
```

What are we looking at?

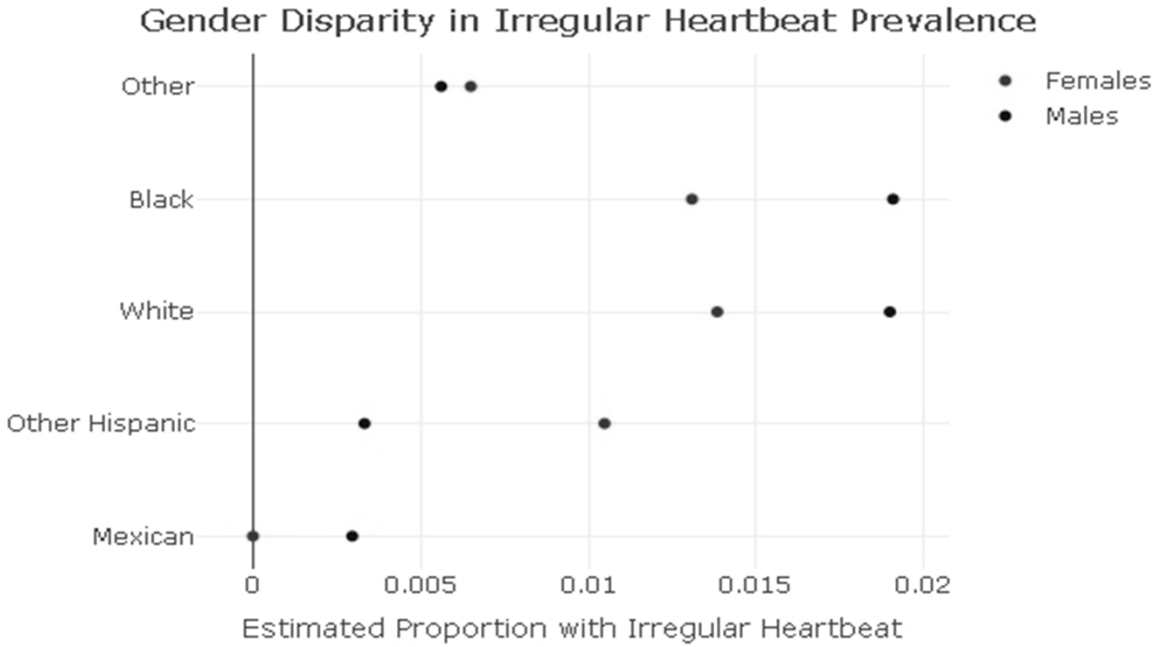
- Next, we create a data set with two unique columns for the male and female estimates, and use the following code to create the dot plot:

```
tempirr2 <- tempirr[tempirr$riagendr == "Male", c(2,3)]
tempirr2$irregularF <- tempirr[tempirr$riagendr == "Female", 3]


p <- plot_ly(tempirr2, x = ~irregularF, y = ~RIDRETH1, name = "Females", type =
"scatter", mode = "markers", marker = list(color = "red")) %>%
  add_trace(x = ~irregular, y = ~RIDRETH1, name = "Males", type = 'scatter', mode =
"markers", marker = list(color = "blue")) %>%
  layout(
    title = "Gender Disparity in Irregular Heartbeat Prevalence",
    xaxis = list(title = "Estimated Proportion with Irregular Heartbeat"),
    yaxis = list(title = ""),
    margin = list(l = 100)
  )
p # show the plot called p in RStudio
```

What are we looking at?

- The result is an **interactive visualization** of the disparities; `plotly` lets you interact with all values being plotted in RStudio!

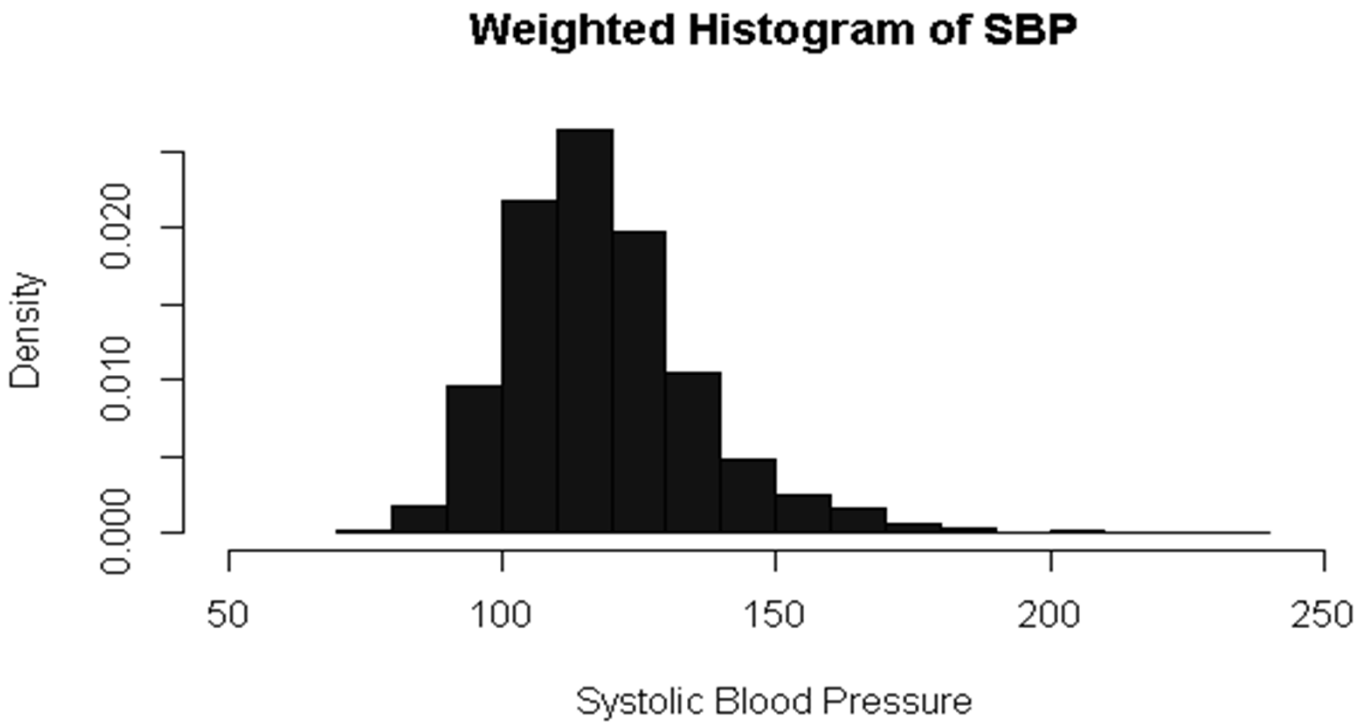


What are we looking at?

- Plots of survey data need to recognize the possibility that sampled observations have different survey weights
- We wish to visualize what the **population** (not the sample!) looks like by employing the survey weights!
- Consider a weighted histogram for a continuous NHANES variable, systolic blood pressure (note the use of **color** and **clearly labeled axes**):

```
svyhist(~BPXSY1, nhanes.dsgn, main="Weighted Histogram of SBP", col =
"blue", xlab = "Systolic Blood Pressure", xlim = c(50,250))
```

What are we looking at?



When were the data collected?

- In all panel surveys, we need to provide clear indications of change in selected survey outcomes over time
- It also helps to paint a picture of how the change over time can vary from one group (or one individual!) to another
- *Trellis plots* in R can do a nice job of indicating variability in trajectories (among subjects or groups) of certain outcomes over time, from panel surveys
- In this illustration, considering trends in socialization between the ages of 2 and 13 for children with autism, we first examine variance in overall trends among three groups defined by the amount of expressive language at age 2

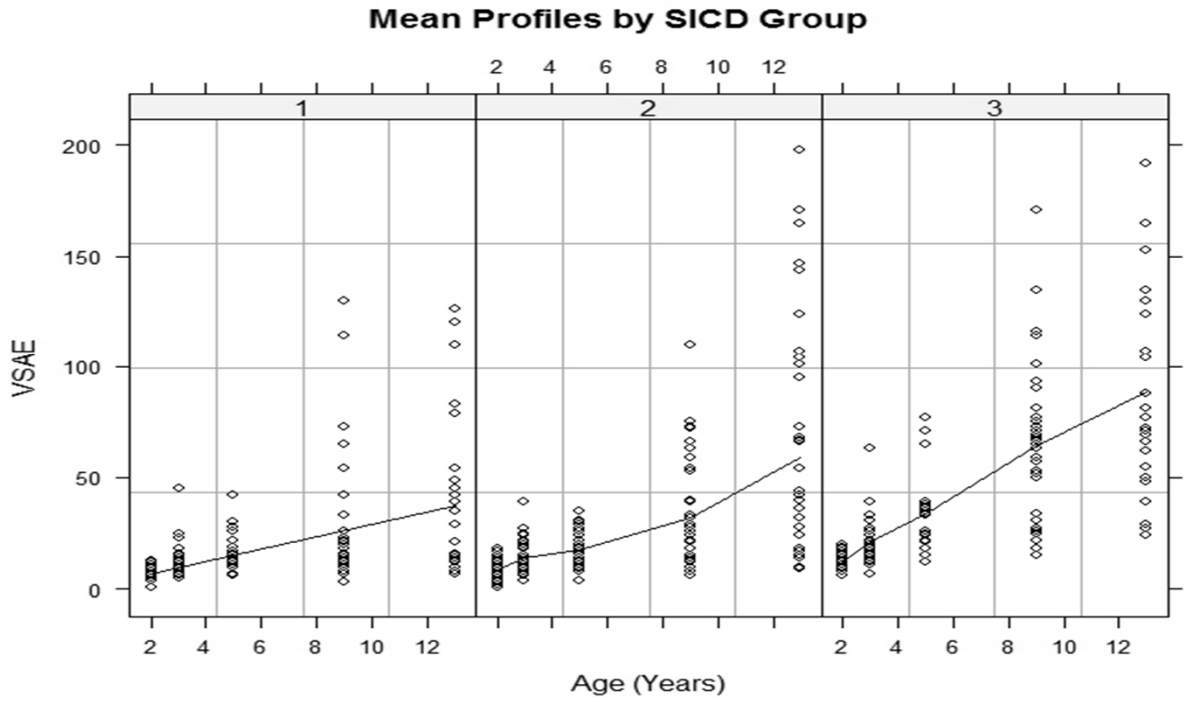
When were the data collected?

```
# load the lattice and nlme libraries (assuming that they are installed)
library(lattice)
trellis.device(color=F)
library(nlme)


# create a groupedData object, with groups based on expressive language (from nlme)
autism.g2 <- groupedData(vsae ~ age | sicdegp.f, order.groups = F, data =
autism.updated) # autism.updated available from previous web link or upon request


# generate the trellis plot
plot(autism.g2, display = "sicdegp", aspect = 2, key = F, xlab = "Age (Years)", ylab
= "VSAE", main = "Mean Profiles by SICD Group")
```

Trends in socialization between ages 2-13

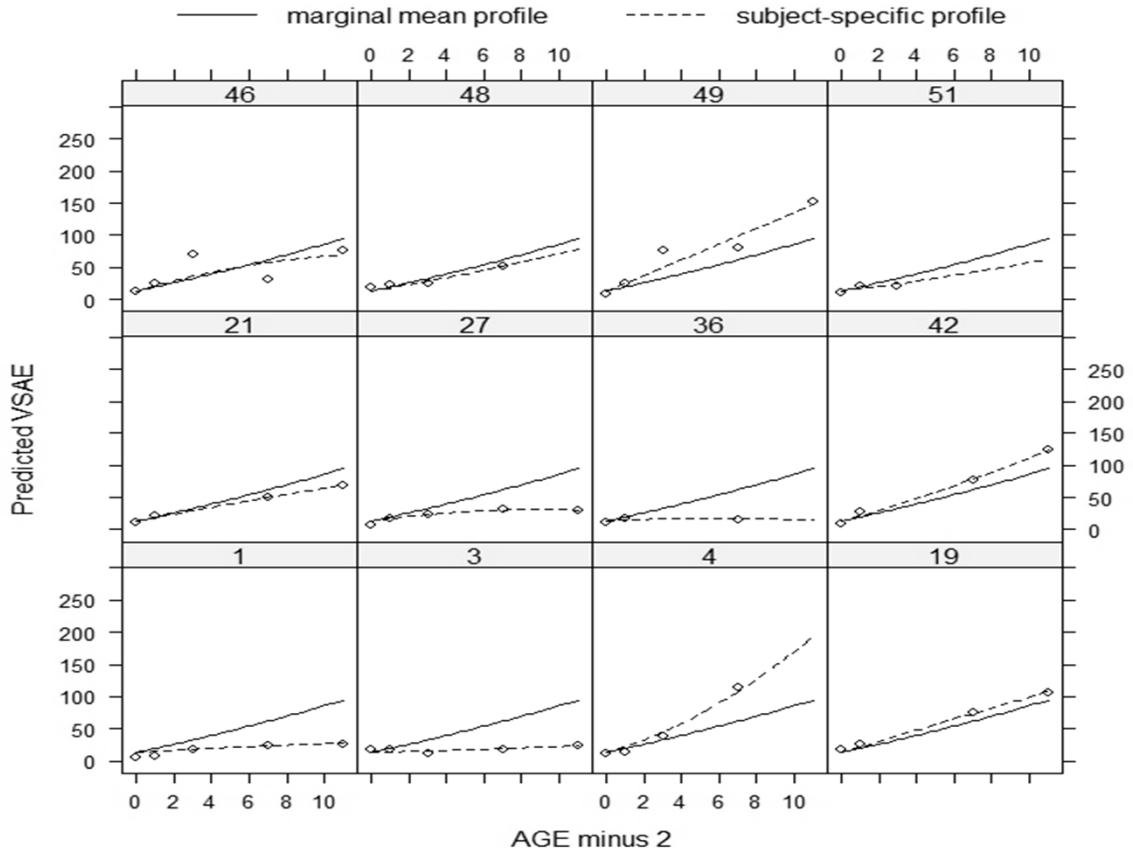


When were the data collected?

- Next, we take the fit of a multilevel model from `lme()`, use the `augPred()` function to generate marginal and subject-specific predictions, and then generate a new trellis plot based on the predictions (separate plots for separate kids; first 12 only!)

```
plot(augPred(model6.3.fit, level = 0:1), layout=c(4,3,1),
xlab="AGE minus 2", ylab="Predicted VSAE",
key = list(lines = list(lty = c(1,2), col = c(1,1),
lwd = c(1,1)), text = list(c("marginal mean profile",
"subject-specific profile")), columns = 2))
```

Subject-specific trends in socialization



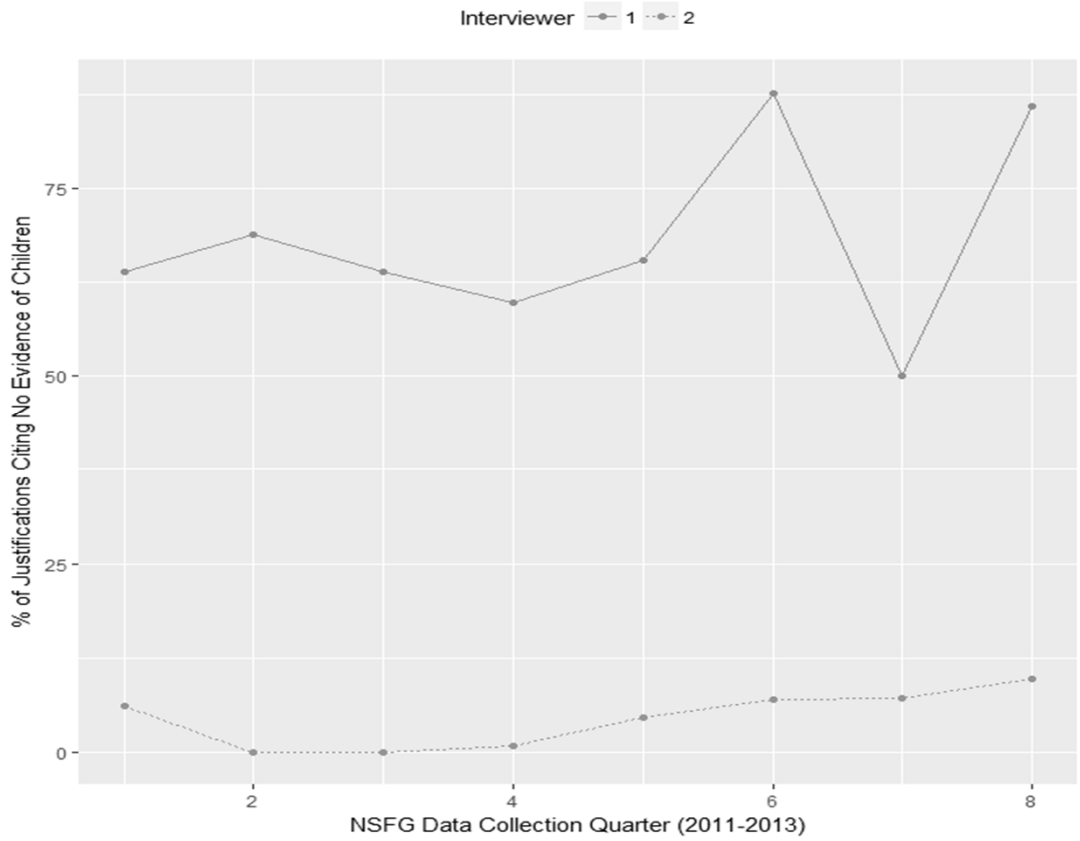
When were the data collected?

- We can also use the ggplot2 package to generate multiple line plots, in this case showing differences in trends over time (quarters) in the % of time an observational strategy is used between groups (interviewers):

```
intdata <- cbind(c(1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2),
c(1,2,3,4,5,6,7,8,1,2,3,4,5,6,7,8),
c(63.96,68.92,64,59.87,65.42,87.76,50,86.08,6.09,0,0,0.83,4.55,6.98,7.14,9.71))
intdata <- data.frame(intdata)
colnames(intdata) <- c("Interviewer","Quarter","Percentage")
intdata$Interviewer <- factor(intdata$Interviewer)


ggplot(data=intdata, aes(x=Quarter, y=Percentage, group=Interviewer)) +
geom_line(aes(linetype=Interviewer, color=Interviewer)) +
geom_point(aes(color=Interviewer)) + ylab("% of Justifications Citing No
Evidence of Children") + xlab("NSFG Data Collection Quarter (2011-2013)") +
theme(legend.position="top")
```

When were the data collected?

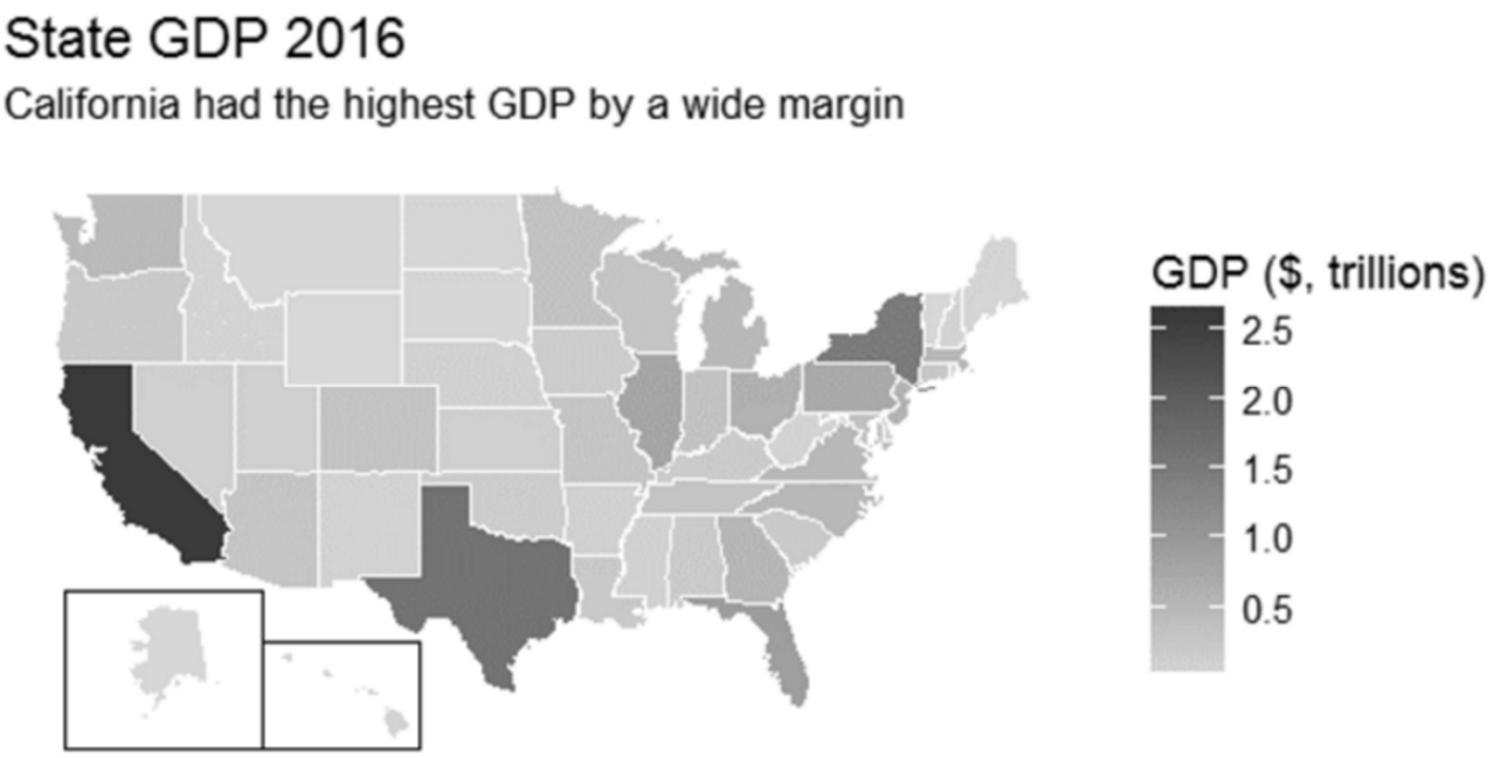


Where are we talking about?

- **Data maps** are an essential tool for visualizing differences between geographic areas in terms of survey measures of interest
- To generate a data map, we need a **shape file** containing the data needed to draw the geographic areas of interest, and a data file containing the information that we wish to show for each area
- We can rely on various combinations of packages within R (including `ggplot2`) to generate these types of data maps
- One useful example, requiring four packages and about 20 lines of code, can be found at the following web site:

https://medium.com/@NickDoesData/visualizing-geographic-data-in-r-fb2e0f5b59c5

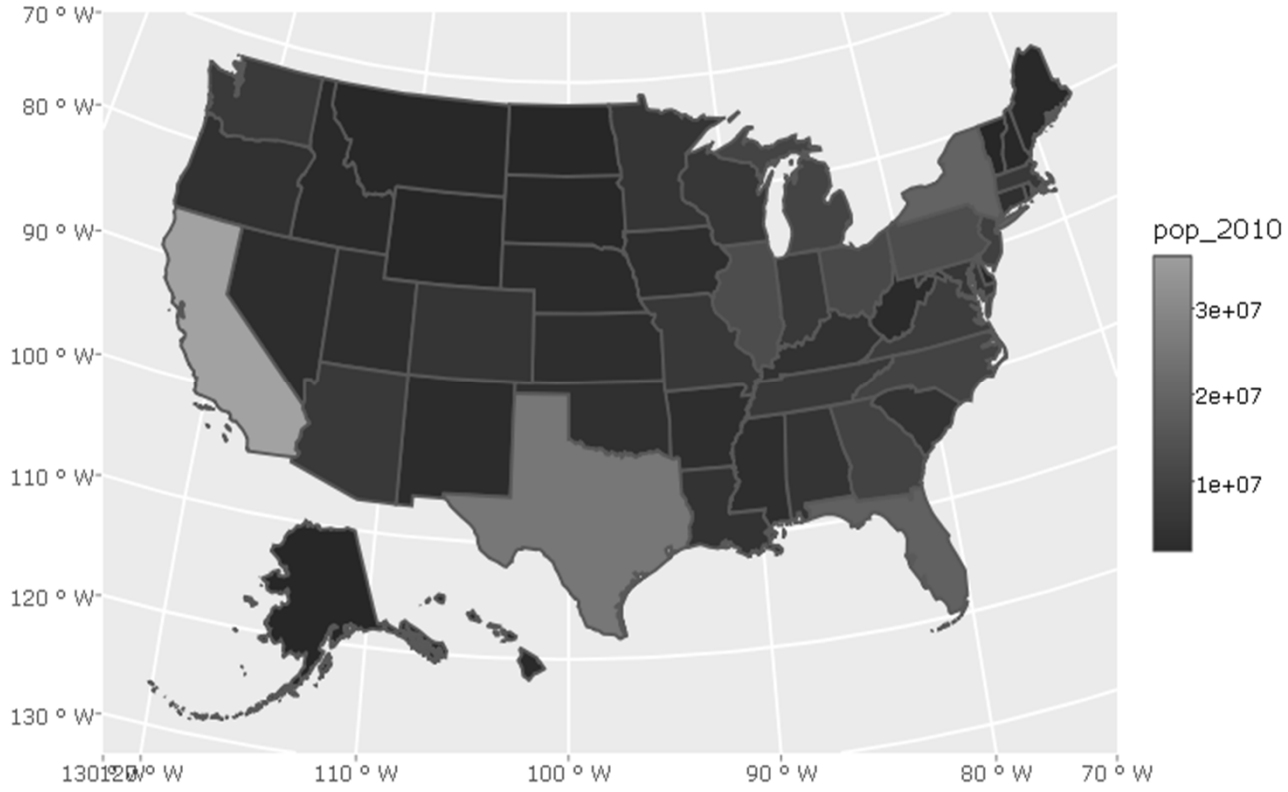
Where are we talking about?



Where are we talking about?

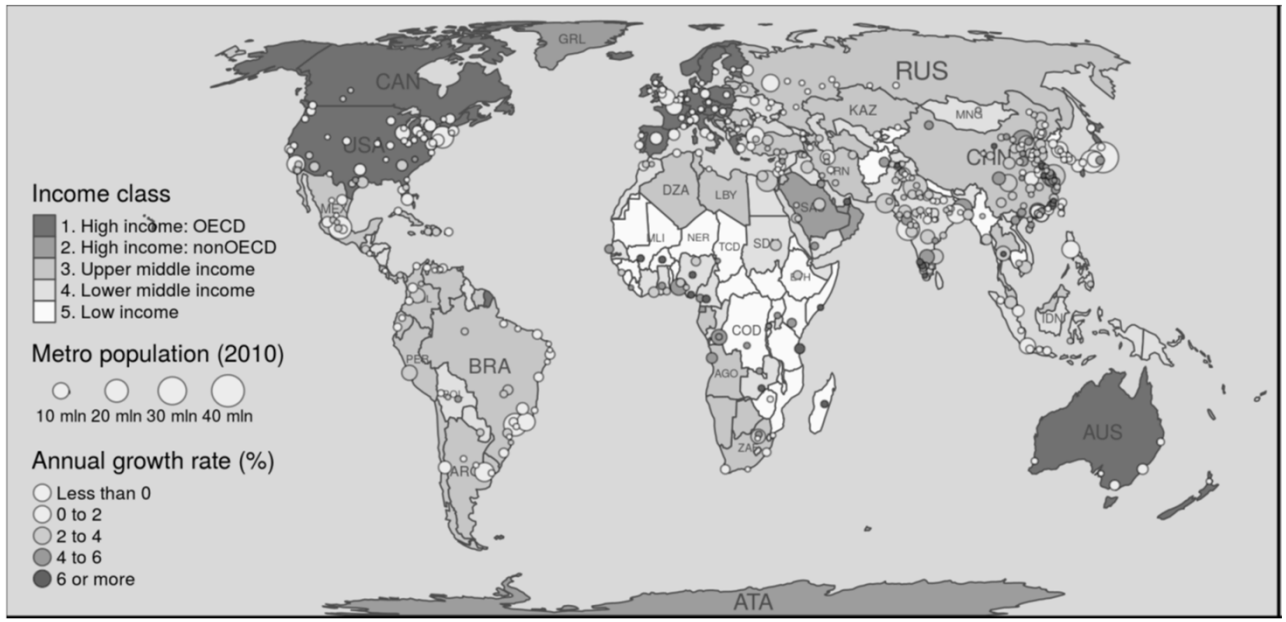
- The `plotly` package is also excellent for generating interactive maps using RStudio: see https://www.r-bloggers.com/plotly-4-7-0-now-on-cran/ for details
- This workshop by Carson Sievert, who is a true master of data mapping in R, provides an excellent overview of how to generate interactive data maps using `ggplot2` and `plotly`
- Here is one more great example from the workshop using `plotly`, which again can be generated with just a couple lines of code:

Where are we talking about?



Where are we talking about?

- Another useful R package that is *brand new* is the `tmap` package: https://www.jstatsoft.org/article/view/v084i06
- Here is an example data map that can be created using this package, given the necessary shape and data files, using the same "layered" graphics employed by the `ggplot2` package:



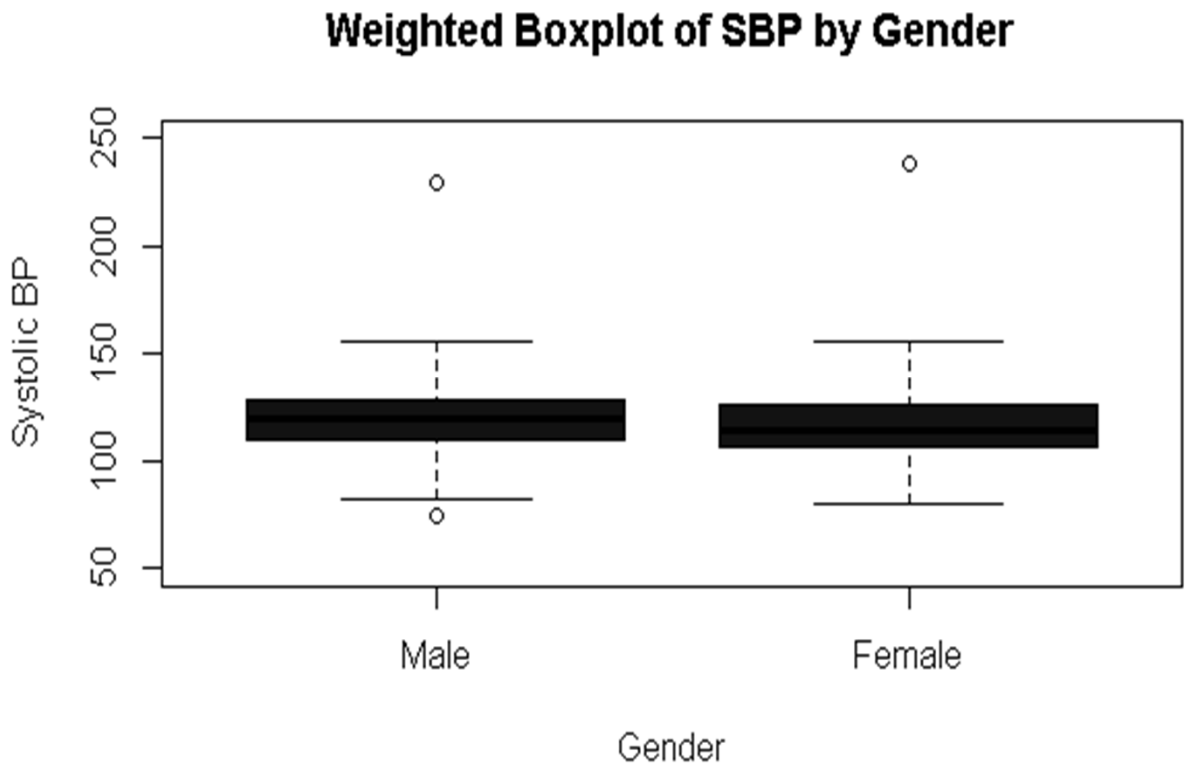
Why were there certain outcomes?

- It is important to keep bivariate and multivariate associations in mind: what are the reasons for particular outcomes?
- **Example:** if a poll is heavily skewed toward older respondents, is age associated with the outcome of interest?
- Consider a weighted side-by-side box plot for comparing distributions on a continuous variable between two groups (from NHANES):

```
svyboxplot(BPXSY1 ~ factor(riagendr), nhanes.dsgn, main="Weighted
Boxplot of SBP by Gender", ylab="Systolic BP", xlab="Gender",
col="blue", ylim = c(50,250))
```

- Note the use of color, value labels, and clearly labeled axes

Why were there certain outcomes?



Why were there certain outcomes?

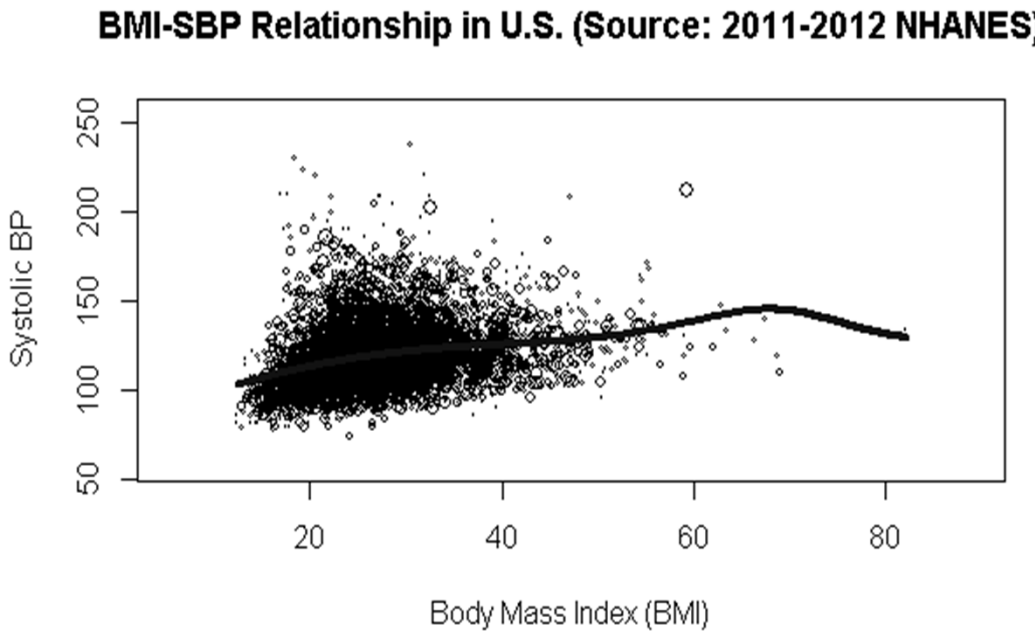
- Next, consider a scatter plot with the individual data points weighted by their respective NHANES survey weights
- We combine this with a weighted smoothing function that visualizes the general (non-parametric) functional relationship between these two variables (Systolic BP and BMI)

```
svyplot(BPXSY1 ~ bmxbmi, design=nhanes.dsgn, xlab="Body Mass Index (BMI)",
ylab="Systolic BP", main="BMI-SBP Relationship in U.S. (Source: 2011-2012
NHANES)")

smoother <- svysmooth(BPXSY1 ~ bmxbmi, design=nhanes.dsgn, bandwidth=10)

lines(smoother, col="blue", lwd=4)
```

Why were there certain outcomes?



The Top 50 visualizations using ggplot2

- This web site is an amazing resource: the Top 50 graphics that one can generate using the `ggplot2` package, with working code!

http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html

- A potential issue: it is not yet entirely straightforward to apply survey weights within `ggplot2` (we've seen some examples…)

- A good start in this direction: the `srvyr` package (https://cran.r-project.org/web/packages/srvyr/vignettes/srvyr-vs-survey.html)

The R Graph Gallery

- This web site is another tremendous resource, allowing readers to click on a given image and see what code was used to generate it:

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

- I would highly recommend checking this web site out if you are looking for some inspiration or have a particular visualization in mind, and have no idea how to get started in terms of the necessary code!

Other Cool Tips and Tricks for #DataViz in R

- Highlighting lines and points in complex graphs using `gghighlight`: https://t.co/SPVeKw0UdN
- Use images for the labels of x-axis ticks: https://www.r-bloggers.com/images-as-x-axis-labels-updated-2/
- A new R package called `default` for quickly changing the default plotting options of various R functions
- Detect and visualize anomalies in time series data

Other Resources for Visualizing Survey Data

- SAS: http://support.sas.com/publishing/vds.pdf
- SPSS: https://www.wiley.com/en-us/SPSS+Statistics+for+Data+Analysis+and+Visualization-p-9781119003557
- Stata: https://www.stata.com/training/onsite-training/courses/data-visualization-in-stata/
- **In my completely unbiased opinion, no other software really offers the flexibility for visualizing survey data that R offers!**
- Some excellent new books on practical #DataViz in R:
  - http://socviz.co/
  - http://serialmentor.com/dataviz/

Brown University's Seeing Theory

- An outstanding training tool for visualize essential concepts of statistics and probability!
- https://students.brown.edu/seeing-theory/
- Relies on **interactive visualization** to communicate these essential concepts
- Highly recommended for those of us who teach and/or consult on these concepts
- Also a useful presentation tool for explaining sampling variability!

Thank you for attending this webinar!

- Please email me at bwest@umich.edu with any follow-up questions!

Web site: http://www.umich.edu/~bwest

Github: https://github.com/bradytwest