Research and Working with Data in R

Mark M. Fredrickson (mfredric@umich.edu)

Computational Methods in Statistics and Data Science (Stats 406)

Final Paper

Overview

- Goal: perform your own investigation using computational techniques.
- Select a research topic with a clear question.
- Find a relevant data set.
- Use class techniques to describe, visualize, and analyze the data.
- Evaluate statistical properties (operating characteristics) of methods
- Two assignments (see Canvas):
  - First draft: Jul 18
  - Final draft: Aug 4

Structure

- Introduction (1 - 2 pages) – present question and motivation
- Data (1 - 2 pages) – describe, plot data, explaining defficiencies
- Method (1 - 3 pages) – motivate several methodological approaches
- Simulations (2 - 3 pages) – evaluate methods for their operating characteristics
- Analysis (3 - 5 pages) – apply best methods to data, interpret results
- Discussion (1 - 2 pages) – summarize, impact, future
- Bibliography/Works Cited
- Supplemental Materials

Total length: 10 - 15 pages

Other Requirements

- All plots and tables should have clear labels (axes, title, legend) and a brief caption.
- At least one plot or table in each of the data, simulations, and analysis sections.
- There should be at least three citations in MLA format (explaining the data source, motivation for the research topic, information on the method)
- No more than one popular press/webpage counts towards minimum
- Citations to course books are accepted
- Clear, well structured writing
- No code in paper

These requirements also given on the Final Project page on Canvas.

First Draft

First draft, Jul 18:

- Clearly set out the research question
- Identify a data set and provide some presentation of data
- Plan for completing rest of paper
- Possibly early results
- Around 3 to 5 pages.
- Goal: get feedback from me.

Opportunity for one-on-one meetings following return of first draft.

Final Draft

Final draft, Aug 4:

- Fully completed paper
- Supplemental materials:
  - All code and data to perform analysis
  - Any additional analysis or simulation not critical to the paper
  - May be structured as one or more .Rmd files and include code.

Second round of one-on-one meetings prior to final due date.

Some Examples

- Does the opioid affect all age groups evenly? Are deaths due particular drugs more common in certain age groups?
  - Data: Accidental drug overdoses in CT, 2012– 2017
  - Methods: Bootstrapped confidence intervals for average age per drug category. Permutation tests of hypothesis that age and category are independent, with power analysis for test statistic selection.
- Are bicycle sharing rental station locations related to the income of a neighborhood?
  - Data: Bike share locations in Brooklyn, NY.
  - Methods: Spatial point process model of locations with parameter that controls dispersion of locations, test statistics that attended to spread and distance of locations, confidence intervals by test inversion.

More Examples

- Can batters predict the next pitch given previous pitches? What factors most closely associate with pitch selection?
  - Data: 100,000 recorded pitches from MLB with 20 features.
  - Methods: Logistic regression with bootstrapped confidence intervals for parameter estimates, simulations to estimate bias and variance under different data generation models
- Do theoretical models relating central black hole size and galaxy shape hold in empirical data? Do elliptical galaxies differ from spiral galaxies in black hole mass?
  - Data: A collection of galaxies classified by shape, with measurements of angular velocity, black hole mass, and other features
  - Methods: Multiplicative model of mass with boostrapped confidence intervals, nearest-neighbor hypothesis test of shape-mass relationship.

Starter Projects

Three projects that include a **data set** and a **related paper** that you can use to jumpstart your project.

- School Shootings
- Bumblee Migration and Climate Change
- 1918 Influenza Epidemic

You still need to **develop your own research question**.

Rubric (1/2)

Out of 80 total points:

- Intro (10): Clear research question, one or more citations to motivate research, summary of subsequent sections of paper.
- Data (10): Data source cited, several plots showing key relationships, discussion of missing data (including if there is none).
- Methods (12): Description of least three class methods, clear explanation of model and assumptions, explain techniques and algorithms to someone without STATS406 background, citation for each method (class notes and textbooks acceptable).

Rubric (2/2)

- Simulations (14): Connect relevant operating characteristics to methods, Monte Carlo investigation of operating characteristics, interpretation of results (including selection of statistics).
- Analysis (14): Apply each method to data, interpret results of tests/estimators/prediction with emphasis on research question.
- Discussion (10): Summarize research question and main findings, connect different methods into cohesive whole, present short comings of methods/data, offer suggestions for future research.
- Overall/Formatting/Labels/Supp. Materials (10): Include code and data in supplemental material, citation in understandable format and bibliography included, proper labels/captions for tables and graphs, no R code in paper body.

Mini Research Project

Migration patterns

> *The Midwest is booming, but not where you might think. Kansas City, Minneapolis, Indianapolis, Columbus, Grand Rapids, and Des Moines are the fastest-growing cities in the Midwest. . . . The coasts' loss ended up, to some extent, as Indianapolis, Minneapolis, Des Moines, and Columbus' gains, reflecting a growing flight from what are increasingly gated cities, affordable only to the affluent, the subsidized (students), and those older residents who bought when the buying was good. High housing costs – sometimes three times higher adjusted for income compared to the rising Midwest cities – make attaining homeownership all but impossible. Joel Kotkin, The Daily Beast, 2018-04-27*

How do people choose where they move? What factors are important? Can we model the growth of towns and cities?

A very simple model of migration: move where your friends are

Let's consider a **stochastic process** that describes the population of each of k cities.

At time $t = 0$, each city has population $p_i = 1$. Then at each time period, a person elects to move to a city with probability

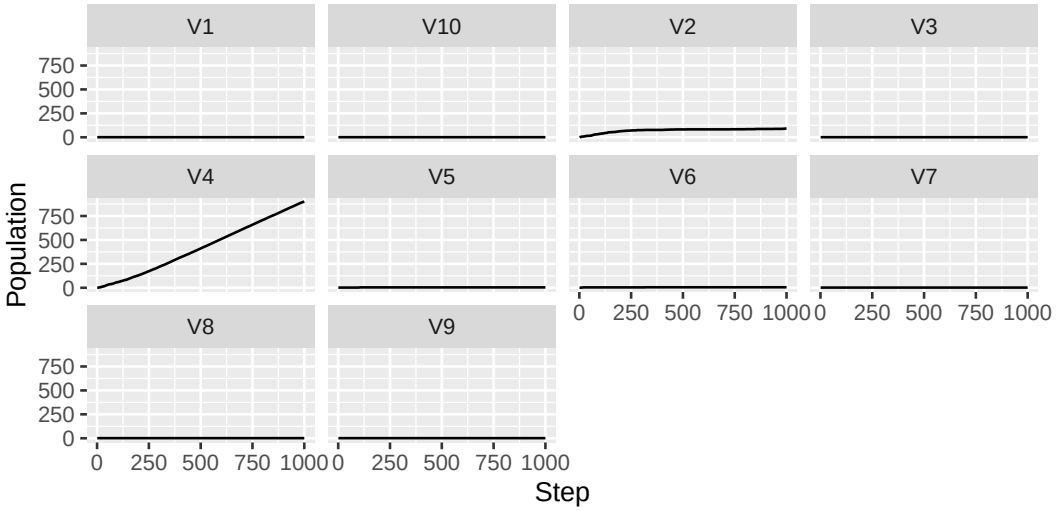
$$P(C(t) = i) = \frac{p_i(t-1)^2}{\sum_{j=1}^{k} p_j(t-1)^2}$$

Then the appropriate city's population is updated, and the process continues.

R implementation

```
> run_pop_process <- function(cities, steps, start = 1) {
+   pops <- matrix(1, nrow = steps, ncol = cities)
+   for (i in 2:steps) {
+     prev <- pops[i - 1, ]
+     ps <- prev^2 / sum(prev^2)
+     which_city <- sample(1:cities, size = 1, prob = ps)
+     prev[which_city] <- prev[which_city] + 1
+     pops[i, ] <- prev
+   }
+   return(pops)
+ }
> pops <- run_pop_process(10, 1000)
```

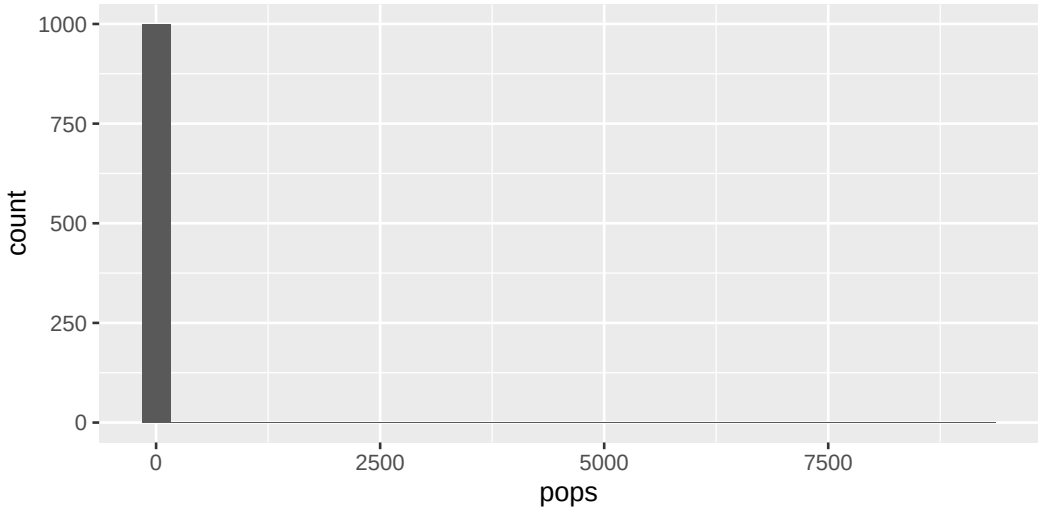
Plotting size v. time



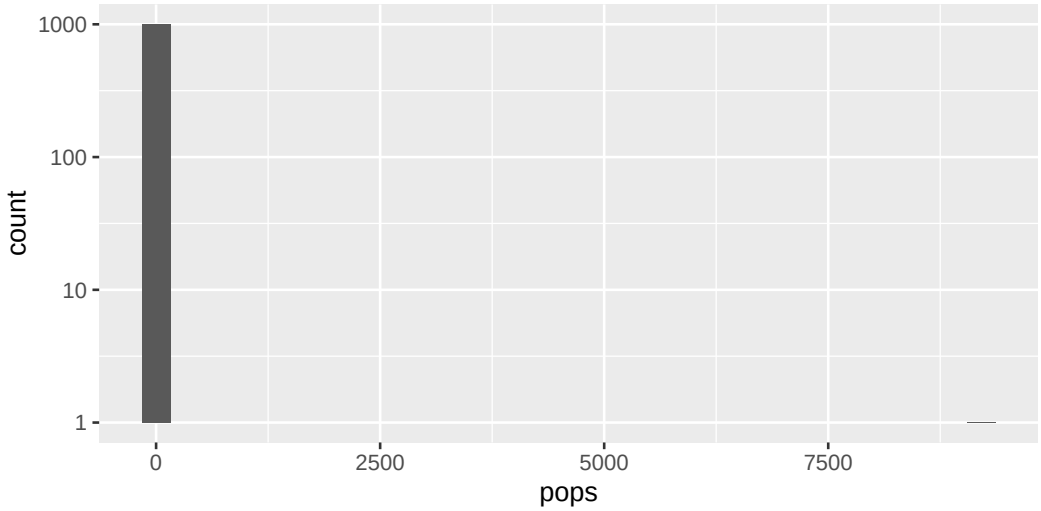
Much larger number of cites

```
> # just get the final value
> big_pops <- run_pop_process(1000, 10000)[10000, ]
> range(big_pops)

[1]    1 9207
```

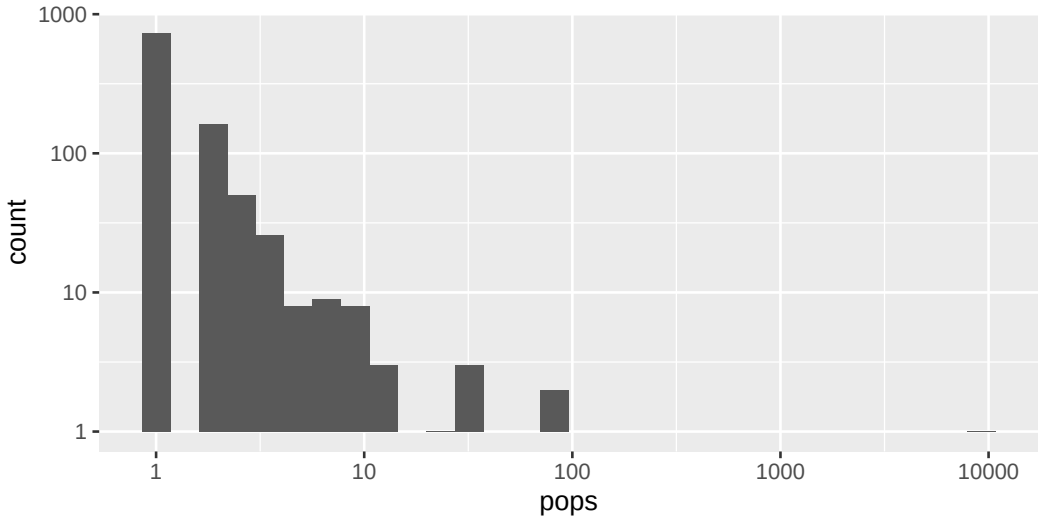
Histogram



Histogram: log y-axis



Histogram: log on both



Power law distributions

A **power law distribution** is defined such that

$$\log(f(x)) = -a \log x + c$$

In other words, the density or mass function is **linear on the log-log scale**.

Perhaps, if cities grow like our simple simulation, a power law distribution would be a good model?

Why study power law distributions?

According to M. E. J. Newman

> *Power-law distributions occur in an extraordinarily diverse range of phenomena. In addition to city populations, the sizes of earthquakes, moon craters, solar flares, computer files and wars, the frequency of use of words in any human language, the frequency of occurrence of personal names in most cultures, the numbers of papers scientists write, the number of citations received by papers, the number of hits on web pages, the sales of books, music recordings and almost every other branded commodity, the numbers of species in biological taxa, people's annual incomes and a host of other variables all follow power-law distributions (Newman, 2005).*

The Pareto Distribution

An example of a power law distribution is the **Pareto distribution**:

$$F_X(x) = 1 - \left(\frac{\beta}{x}\right)^r, \quad x \geq \beta > 0, r > 0$$

To simplify matters, we will assume that $\beta = 1$ (could revisit this assumption later).

Simulations

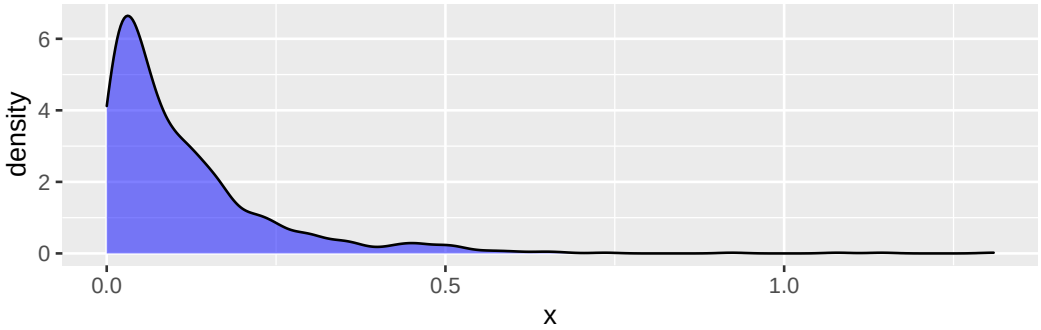
We can generate from a Pareto distribution using a mixture distribution of the form (random parameters):

$$L \sim \text{Gamma}(r, 1), \quad X \sim 1 + \text{Exponential}(L)$$

```
> rpareto <- function(n, r) {
+     ls <- rgamma(n, r, 1)
+     rexp(n, ls)
+ }
```

Example estimated density when $r = 10$

```
> ggplot(data.frame(x = rpareto(1000, 10)), aes(x = x)) +
+     geom_density(fill = "blue", alpha = 0.5)
```



Estimating r

Newman suggests the **maximum likelihood estimator**:

$$\hat{r}_{MLE} = 1 + n \left[\sum_{i=1}^{n} \log(X_i) \right]^{-1}$$

Another option would be to use a methd of moments estimator:

$$E(X) = \frac{r}{r-1} \Rightarrow r_{MoM} = \frac{\bar{X}}{\bar{X} - 1}$$

(Wikipedia, "Pareto Distribution", 2018)

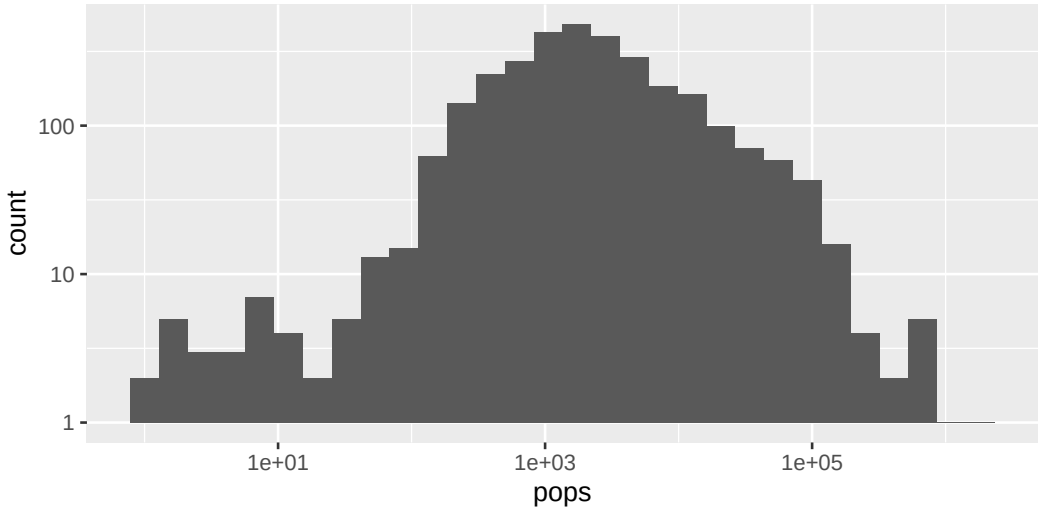
Estimating MSE

```
> r_mle <- function(x) { 1 + 1 / (mean(log(x)))}
> r_mom <- function(x) { mean(x) / (mean(x) - 1)}
> samples <- rerun(1000, rpareto(10, r = 10))
> mean(map_dbl(samples, ~ (r_mle(.x) - 10)^2))

[1] 87.61

> mean(map_dbl(samples, ~ (r_mom(.x) - 10)^2 ))

[1] 102.6
```

Data

```
> # read the data and drop the first (garbage) row
> michigan_complete <- read.csv("sub-est2017_26.csv",
+                               stringsAsFactors = FALSE)[-1, ] %>%
+    mutate(CENSUS2010POP = as.numeric(CENSUS2010POP)) %>%
+    filter(CENSUS2010POP > 0, FUNCSTAT == "A")
>
```

City size distribution



Estimating r

```
> r_mle(na.omit(michigan_complete$CENSUS2010POP))
```

```
[1] 1.131
```

As we saw, the distribution did not look much like our typical Pareto (a mixture of exponentials), and this parameter is **the minimum possible r**.

Further directions

- Investigate fit of Pareto distribution, consider other distributions
- Other data sources, more variables, conditional distributions.
- How robust are the different test statistics to deviations from the assumptions (e.g., $\beta > 1$, data are dependent in some fashion, mixture distributions).
- Further considering why power law distributions are useful (or not! See Clauset et al. (2009)).

Refences

Clauset, A.; Shalizi, C. R. & Newman, M. E. J. (2009) "Power-Law Distributions in Empirical Data." SIAM Review, 51, 661-703

Newman, M. E. J. (2005) "Power laws, Pareto distributions and Zipf's law." Contemporary Physics, 46:5, 323-351

Wikipedia (2018) "Pareto distribution." Retrieved 2018-10-02. `https://en.wikipedia.org/wiki/Pareto_distribution`

Working with real data

Data Formats

Generally, we will find two types of data file formats in the wild:

- Proprietary: .xls or .xlsx (Excel), .dta (Stata), .sqlite (SQLite)
- Delimited: tab, comma, other

Proprietary

Support for these varies. For example, support for .xlsx files has been hit or miss in the past.

Best option: load in other software, convert to a delimited format.

The `foreign` library can read some formats (SPSS, Stata, SAS).

Delimited

A **delimited** data file is a **text file** where the columns are separated by special characters:

```
1, "some text", 8, 2.30
2, "other text" ,, 9.88
3, "text, with comma", 0, -5.23
```

Common delimiters include commas and tabs:

```
read.csv
read.table
```

Also the more generic `read.delim`.

Loading Data in R

Data will often come in compressed formats. For example, .zip or .gz. Your first step is to decompress it.

Then load it into a variable using a `read` function.

You may need to convert some columns to different data types (using `as.SOMETHING`)

Saving Data in R

R has its own proprietary format, .rda.

You can save not only single tables, but collections of information, functions, variables.

```
> k <- 100000
> myrands <- rnorm(k)
> save(file = "example.rda", k, myrands)
```

Loading saved files

```
> rm(k) ; rm(myrands)
> load("example.rda")
> print(k)

[1] 1e+05
```

Basic Data Cleaning

Real data often has missing values, bad values.

It is useful to investigate to see if there missing values or things that can't happen.

```
> summary(somedata)
```

```
       x1                x2
 Min.   :-2.587   Length:100
 1st Qu.:-0.621   Class :character
 Median : 0.083   Mode  :character
 Mean   :-0.014
 3rd Qu.: 0.562
 Max.   : 2.958
 NA's   :5
```

We could delete things with missing, but it might be better to **impute** a value

```
> somedata$x1[is.na(somedata$x1)] <- mean(somedata$x1, na.rm = TRUE)
> summary(somedata)
```

```
       x1                 x2
 Min.   :-2.5868   Length:100
 1st Qu.:-0.5186   Class :character
 Median :-0.0136   Mode  :character
 Mean   :-0.0136
 3rd Qu.: 0.5593
 Max.   : 2.9579
```

Example: Loading data from the National Health And Nutrition Examination Survey

The **National Health And Nutrition Examination Survey (NHANES)** provides survey data on the dietary and health habits of people in the United States.

For many years, **low dose aspirin** was thought to be beneficial for those at risk of **heart disease**.

I downloaded data on **aspirin use** and **blood pressure exams** for survey participants.

Opening data

There are two data files, both in XPT format (a SAS format). rseek.org suggested the haven package:

```
> library(haven)
> aspirin <- read_xpt("./RXQASA_H.XPT")
> dim(aspirin)
```

```
[1] 3815    8
```

```
> bp <- read_xpt("./BPX_H.XPT")
> dim(bp)
```

```
[1] 9813   23
```

Coding “taking aspirin” variable

```
> ### Questions:
> # RXQ515 - Followed (doctor's) advice, took low-dose aspirin?
> # RXQ520 - Taking low-dose aspirin on your own?
>
> eq1 <- function(x) {
+     tmp <- x == 1
+     tmp[is.na(tmp)] <- FALSE
+     return(tmp)
+ }
> aspirin$taking_aspirin <- eq1(aspirin$RXQ515) | eq1(aspirin$RXQ520)
```

Aggregating multiple blood pressure readings

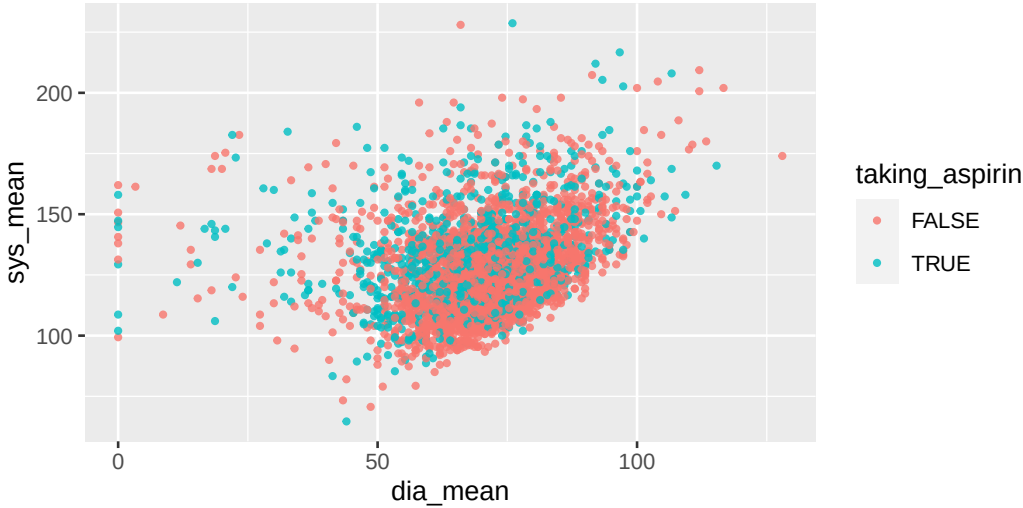
```
> bp$sys_mean <- rowMeans(bp[, c("BPXSY1", "BPXSY2", "BPXSY3", "BPXSY4")],
> bp$dia_mean <- rowMeans(bp[, c("BPXDI1", "BPXDI2", "BPXDI3", "BPXDI4")],
```

Combining

The `dplyr` (part of `tidyverse`) has several methods for **joining tables**:

- `inner_join` matches the tables on a **common key**, discard any entries that do not have a match (multiple matches possible)
- `left_join` keeps everything in the first table, even if no match (`NA` values for unmatched)
- `full_join` keeps everything in both tables, matching were it can (again, `NA` for unmatched)

```
> combined <- inner_join(aspirin, bp, "SEQN") # common "sequence number" I
```


sys_mean
200
150
100
0
50
100
dia_mean
taking_aspirin
FALSE
TRUE

Final Thoughts

- Document, document, document: use scripts/RMarkdown documents to record your changes to data
- Investigate outliers, missing values, strange patterns. Does -9 make sense for count data?
- Use `rseek.org` with your file extension to find packages.
- Open in other software and convert to CSV files.