Welcome to AEM 2850 / 5850!

Week 1

AEM 2850 / 5850 : R for Business Analytics
Cornell Dyson
Spring 2023

Acknowledgements: Andrew Heiss, Claus Wilke, Laurent Bergé

Plan for today

Why take R for Business Analytics?

Summary of key class details

Teaser example

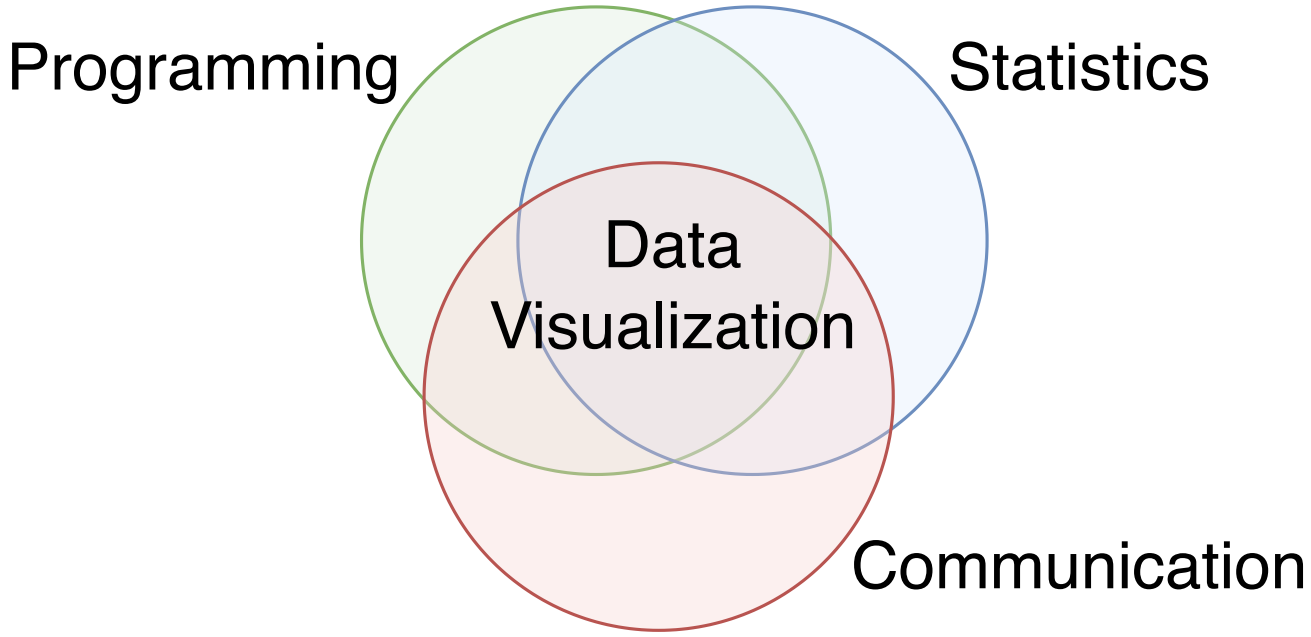
Just show me the data!

What makes a great visualization?

Self-introductions (time permitting)

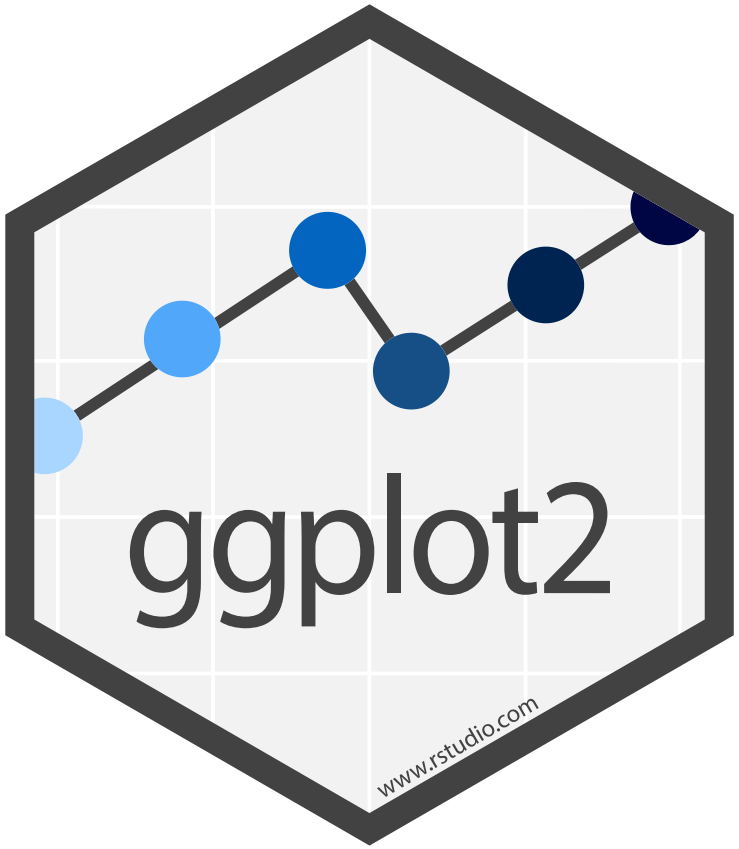
Why take R for Business Analytics?

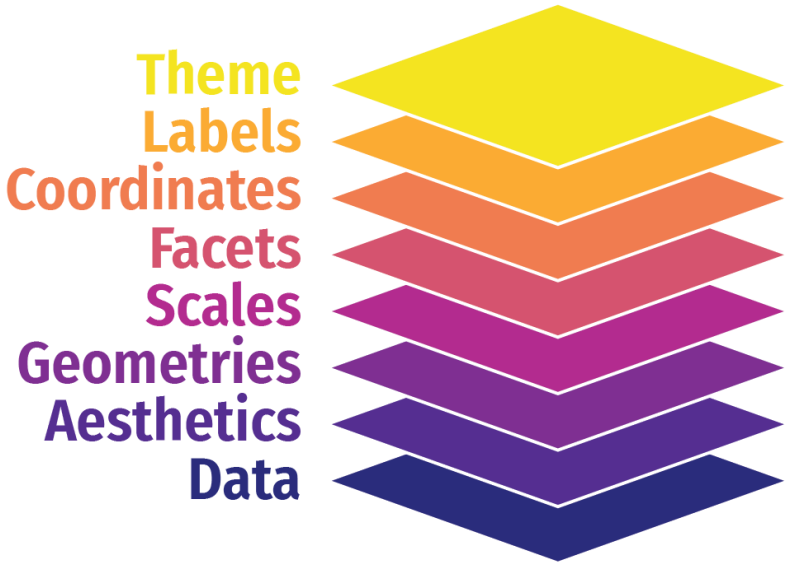
Why take R for Business Analytics?



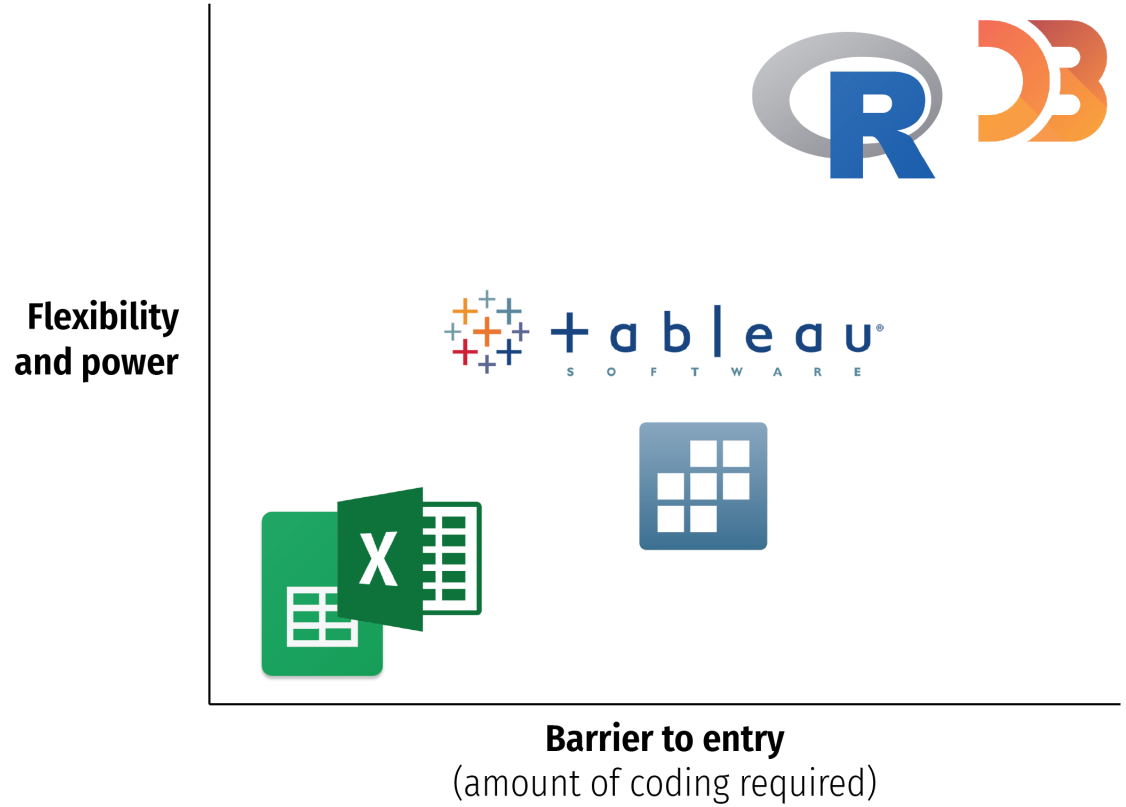
Why R for Business Analytics?

Why R for Data Visualization?





Why R for Data Visualization?



Why R for Life?

Practical tool that could help you get a job and then do said job

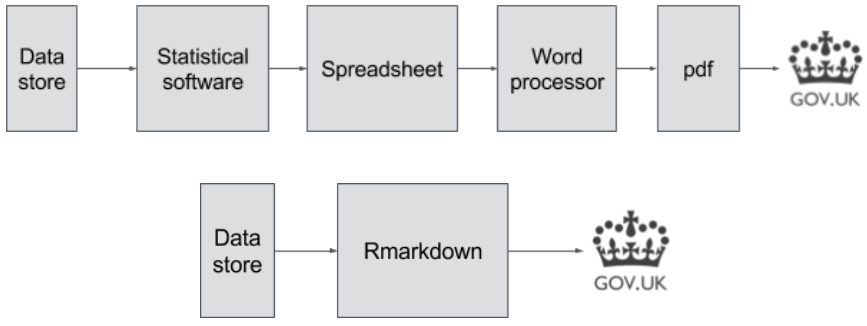
3.1.2 Data Visualization

We use ggplot2 as our main package to create ad-hoc exploratory graphics as well as polished-looking customized visualizations. When combined with tools to clean and transform data, ggplot2 allows analysts to quickly translate insights into high quality, compelling visualizations. In addition to the static graphics of `ggplot2`, we often make interactive visualizations or dashboards using R packages such as `plotly` (Sievert et al. 2017), `leaflet` (Cheng et al. 2017), `dygraphs` (Vanderkam et al. 2017), `DiagrammeR` (Sveidqvist et al. 2017), and `shiny` (Chang et al. 2017).

3.1.3 Reproducible Research

At Airbnb, all R analyses are documented in `rmarkdown`, where code and visualizations are combined within a single written report. Posts are carefully reviewed by experts in the content area and techniques used, both in terms of methodologies and code style, before publishing and sharing with the business partners. The peer review process is

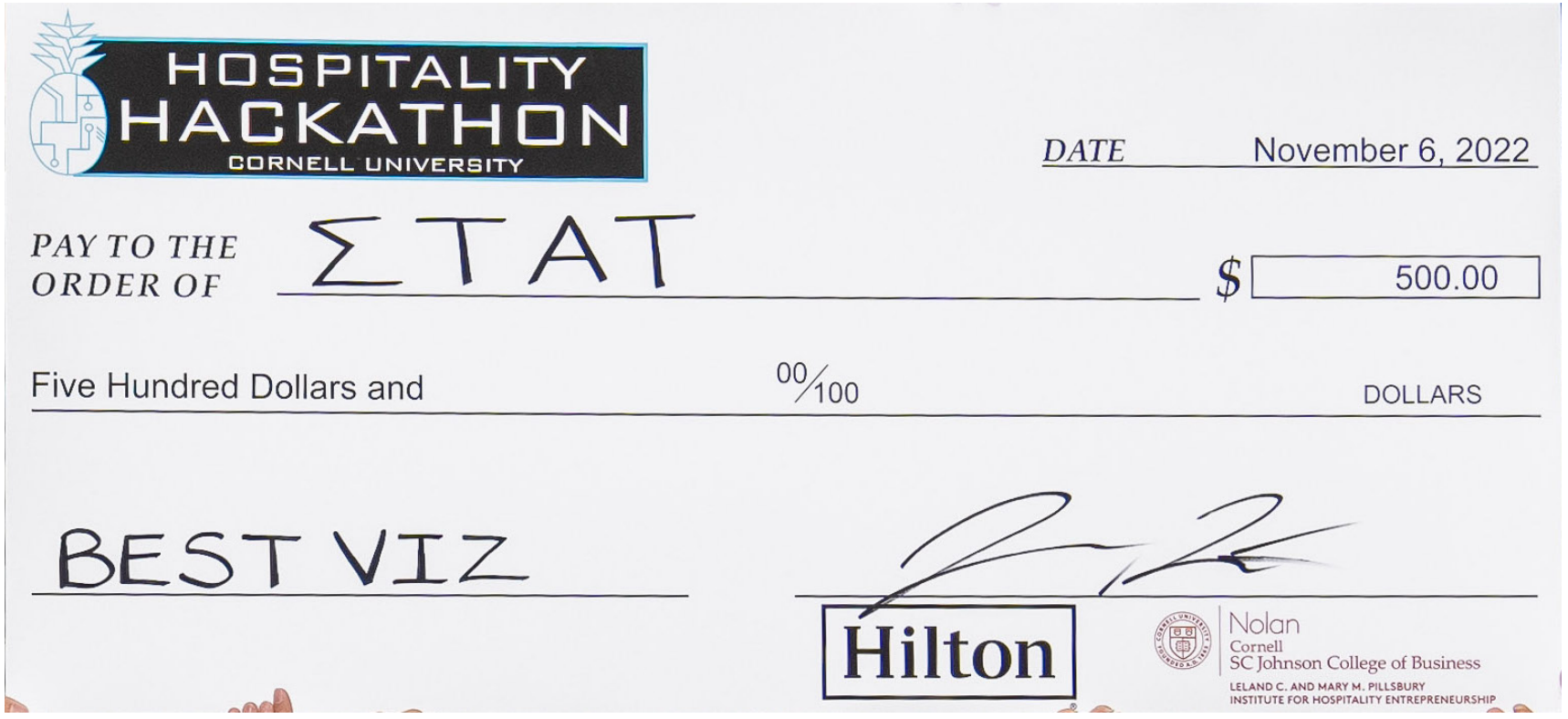
Airbnb, ggplot, and rmarkdown



The UK's reproducible analysis pipeline

Why R for Life?

Practical tool that could help you get a job and then do said job



Or start making money now!

Why R for Life?

Practical tool that could help you get a job and then do said job

Open source

Huge community of users and package developers

Here are a few examples of other things you can do using R:

- Make slides like the ones you're looking at right now
- Build websites like our course site
- Write books like R for Data Science
- Make interactive web apps
- Much, much more

Skills you develop in this course can also be used for other programming languages

Class details

Preface

1. Your success in this class is important to me
2. This course is a work in progress
3. Get the semester off to a good start: **read the syllabus**!

A bit about me

- Prof. Todd Gerarden
- Economist
- Joined Cornell in 2018
- Interested in:
  - Renewable energy
  - Innovation in energy tech
  - Working with data

A bit about our TAs

Graduate TA

Hui Zhou

Undergraduate TAs

Sophie McComb

Jonathan Gotian

We will post office hours and contact information on the course site and canvas

A bit about you

Do you have any programming experience? (None is required or even expected!)

What programming language(s) have you used before?

- R
- Python
- SQL
- VBA
- MATLAB
- Stata
- Other

First course assignment will be to fill out a survey to tell us more about you

We'll also do brief self-introductions at the end of class today if time permits

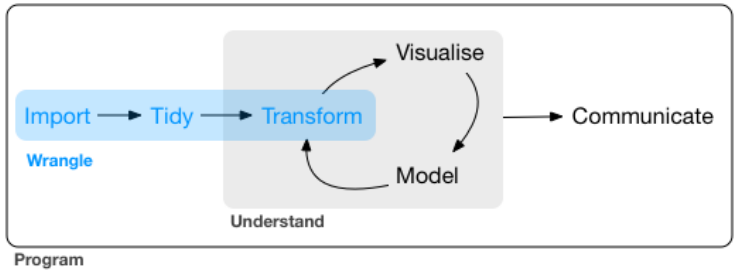
Course objectives

1. Develop basic proficiency in R programming
2. Understand data structures and manipulation
3. Describe effective techniques for data visualization and communication
4. Construct effective data visualizations
5. Utilize course concepts and tools for business applications

Plan for the semester

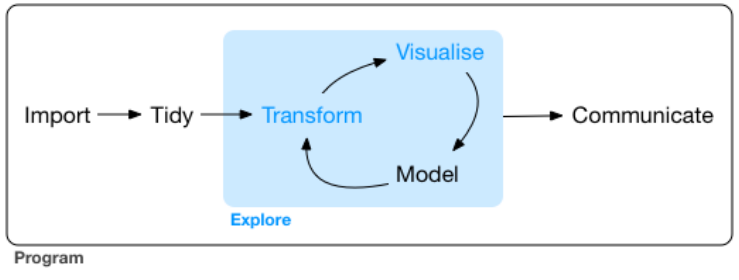
Programming Foundations

`R`, `RStudio`, `R Markdown` / `Quarto`, the `tidyverse`



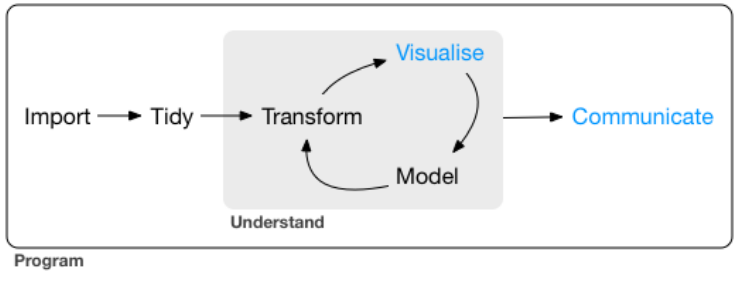
Data Visualization Foundations

the grammar of graphics, `ggplot2`



Special Topics

annotations, time, space, etc.



Plan for each week

We will follow the same general process each week:

- Do readings listed on the course site before Tuesday (example: Week 1)
- **Tuesday:** come to class, where we will discuss material for that week's topic
- **Thursday:** come to class, where we will work through hands-on examples
- Work on the lab before the next Tuesday's class, attending office hours as needed
- **The following Monday:** submit lab on canvas by 11:59pm (starting with Week 1)

Assignments

- **Labs** are short weekly homework assignments that require you to practice programming
- **Prelims** are intended to assess programming and data visualization proficiency
- The **group project** is intended to synthesize and reinforce skills in real-world applications
- **Class participation** is the best way to learn the material, attendance is expected

Students in AEM 5850 complete extra assignments

Assignment	Percent
Labs	35%
Prelim 1	20%
Prelim 2	20%
Group project	20%
Class participation	5%
Total	100%

Contacting us

Office hours:

- TAs: TBD
- Tuesdays 11:00am - 12:00pm: Prof. Gerarden in Warren 466
- Other times by appointment: Prof. Gerarden, at aem2850.youcanbook.me

Email:

You can also reach us by email. The best approach is to email both me and our grad TA Hui Zhou at the same time. You can do that with one click here. Please read the syllabus for tips on how to make the most of email.

Course websites

Site for accessing course materials: (↓)

aem2850.toddgerarden.com

Site for submitting work: (↑)

canvas.cornell.edu/courses/50706

- viewing announcements
- viewing grades
- for convenience, you can also view and navigate the course site through canvas (Home, Syllabus)

Sucking

> "The bad news is whenever you're learning a new tool, for a long time you're going to suck. It's going to be very frustrating.
>
> But, the good news is that that is typical, it's something that happens to everyone, and it's only temporary.
>
> Unfortunately, there is no way to go from knowing nothing about a subject to knowing something about a subject and being an expert in it without going through a period of great frustration and much suckiness.
>
> But remember, when you're getting frustrated, that's a good thing, that's temporary, keep pushing through, and in time [it] will become second nature."

Hadley Wickham, author of `ggplot2`, *R for Data Science*, and much more

I *promise* you can succeed in this class. Don't hesitate to get help from me, TAs, office hours, and your peers.

Questions about the class?

Teaser example

How does 2023 compare to 2022 so far?

Go to aem2850.toddgerarden.com/content/01-content

Click the links to download the following files:

- Weather stations in NY
- Weather in NY in 2022
- Weather in NY in 2023

Make a plot that compares the evolution of daily max temps (TMAX) over January in 2022 and 2023

Use any software you like!

Feel free to work in small groups

How does 2023 compare to 2022 so far?

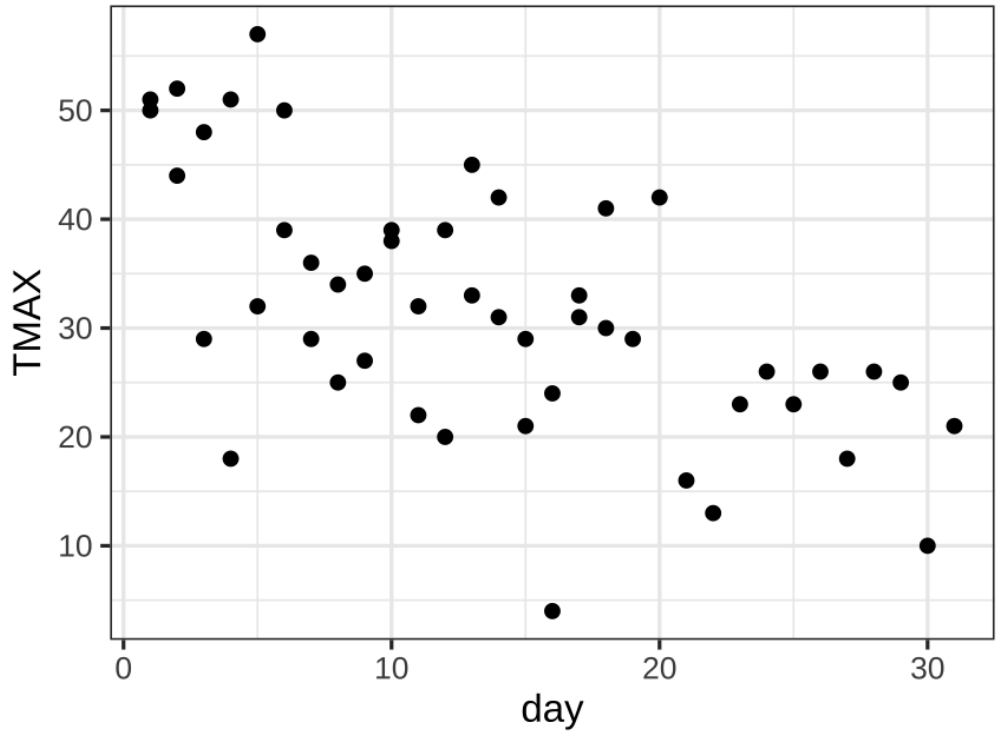
One way to do this in R. First, we'll need to import and prep the data:

```
library(tidyverse); library(lubridate)

# identify the Cornell station
stations <- read_csv("data/01-slides/ny-stations.csv")
cornell <- stations |>
  filter(str_detect(NAME, "CORNELL"))

# read in and bind relevant data
clean_data <- function(y, s, m) {
  str_glue("data/01-slides/ny-weather-", y, ".csv") |>
    read_csv() |>
    inner_join(s, by = "STATION") |>
    mutate(mon = month(DATE),
           day = day(DATE),
           year = year(DATE)) |>
    filter(mon == m)
}
years <- c(2022, 2023)
cornell_temps <- map(years, clean_data, cornell, 1) |>
  bind_rows()
```

How does 2023 compare to 2022 so far?

```
# plot data
cornell_temps |>
  ggplot(aes(x = day,
             y = TMAX)) +
  geom_point() +
  theme_bw()
```

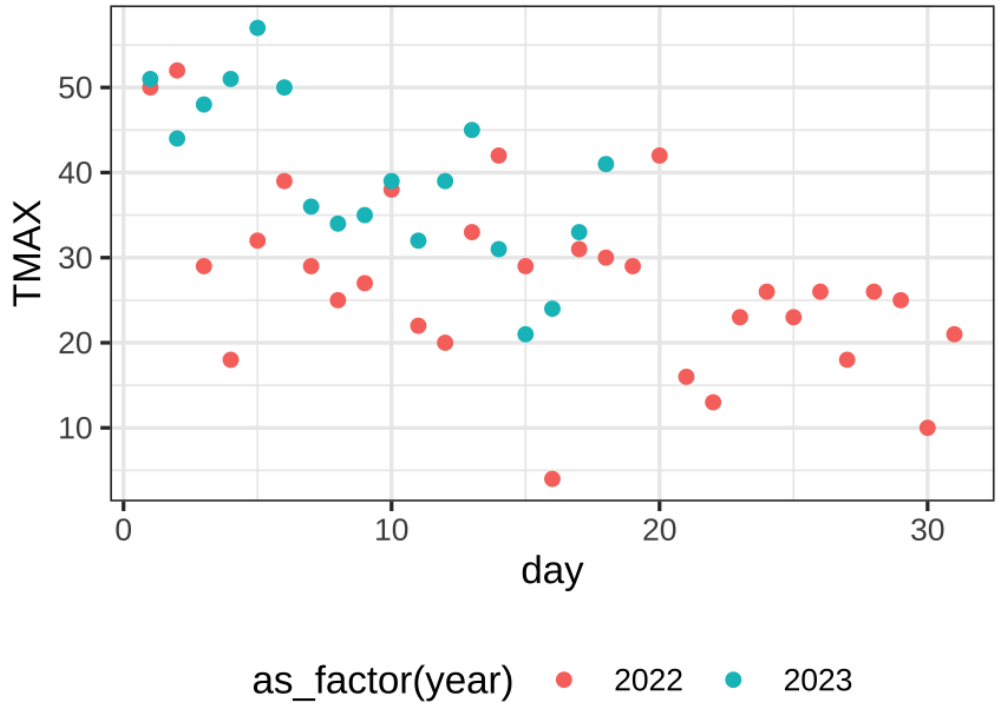
What's wrong with this plot?



How does 2023 compare to 2022 so far?

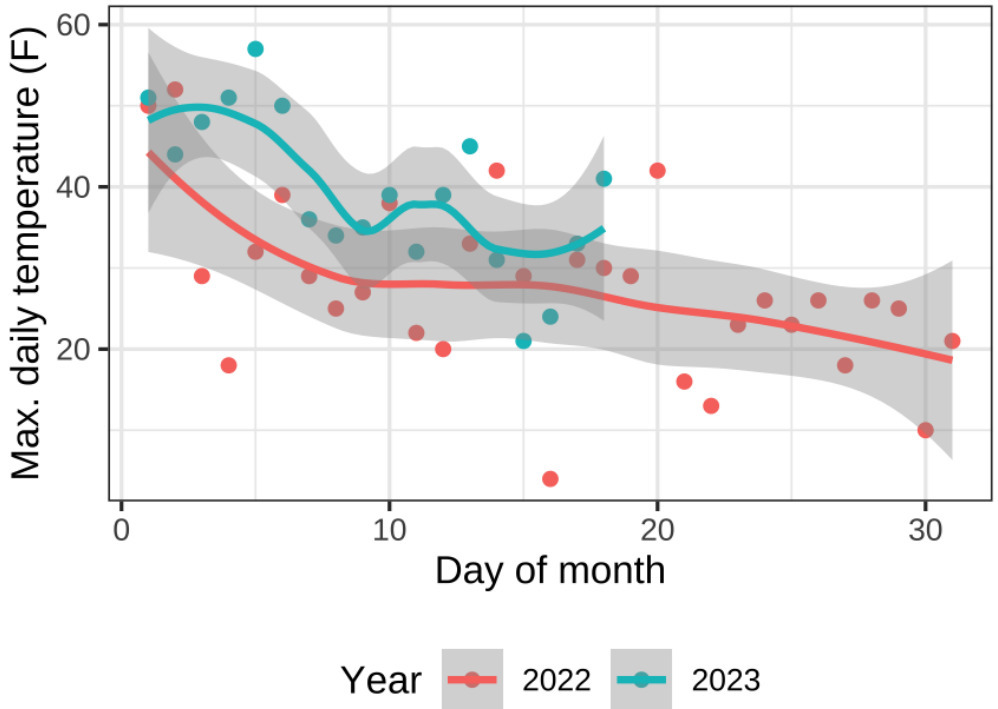
```
# plot data
cornell_temps |>
  ggplot(aes(x = day,
             y = TMAX,
             color = as_factor(year))) +
  geom_point() +
  theme_bw() +
  theme(legend.position = "bottom")
```

What's wrong with this plot?



How does 2023 compare to 2022 so far?

```
# plot data
cornell_temps |>
  ggplot(aes(x = day,
             y = TMAX,
             color = as_factor(year))) +
  geom_point() +
  geom_smooth() +
  theme_bw() +
  theme(legend.position = "bottom") +
  labs(x = "Day of month",
       y = "Max. daily temperature (F)",
       color = "Year")
```



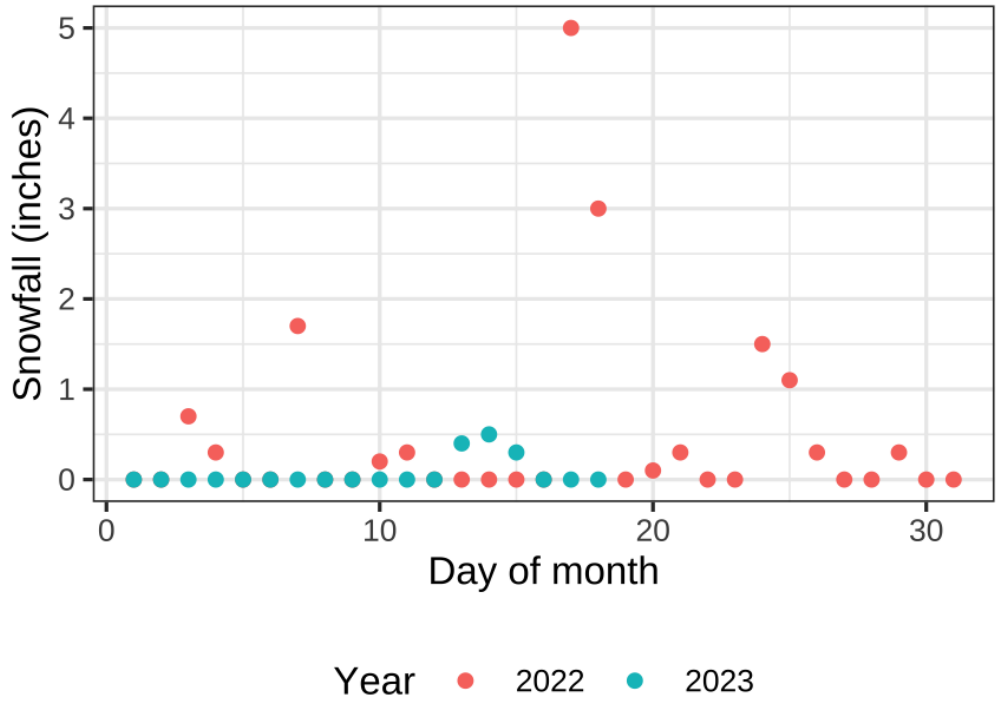
Two advantages of this approach are:

1. we have a script to **reproduce** our work / share our methods with others
2. we could **generalize** and **scale** this much more easily than manual work in excel (for example)

How does 2023 compare to 2022 so far?

For example we can **generalize** this approach to other weather outcomes:

```
# plot data
cornell_temps |>
  ggplot(aes(x = day,
             y = SNOW,
             color = as_factor(year))) +
  geom_point() +
  theme_bw() +
  theme(legend.position = "bottom") +
  labs(x = "Day of month",
       y = "Snowfall (inches)",
       color = "Year")
```



Note: NOAA publishes data with a slight lag so this does not reflect snow since last week

Just show me the data!

Just show me the data!

Data is very powerful, but raw data is not usually enough

```
cornell_temps |>
  group_by(year) |>
  summarize(mean_max = mean(TMAX))
```

```
## # A tibble: 2 × 2
##    year mean_max
##   <dbl>    <dbl>
## 1  2022     27.4
## 2  2023     39.5
```

What's wrong with this calculation?

```
cornell_temps |>
  group_by(day) |>
  filter(n() != 1) |>
  group_by(year) |>
  summarize(mean_max = mean(TMAX))
```

```
## # A tibble: 2 × 2
##    year mean_max
##   <dbl>    <dbl>
## 1  2022     30.6
## 2  2023     39.5
```

Just show me the data!

Here's another example:

```
head(my_data, 10)
```

```
## # A tibble: 10 × 2
##        x     y
##    <dbl> <dbl>
##  1  55.4  97.2
##  2  51.5  96.0
##  3  46.2  94.5
##  4  42.8  91.4
##  5  40.8  88.3
##  6  38.7  84.9
##  7  35.6  79.9
##  8  33.1  77.6
##  9  29.0  74.5
## 10  26.2  71.4
```

```
mean(my_data$x)
```

```
## [1] 54.26327
```

Seems reasonable

```
mean(my_data$y)
```

```
## [1] 47.83225
```

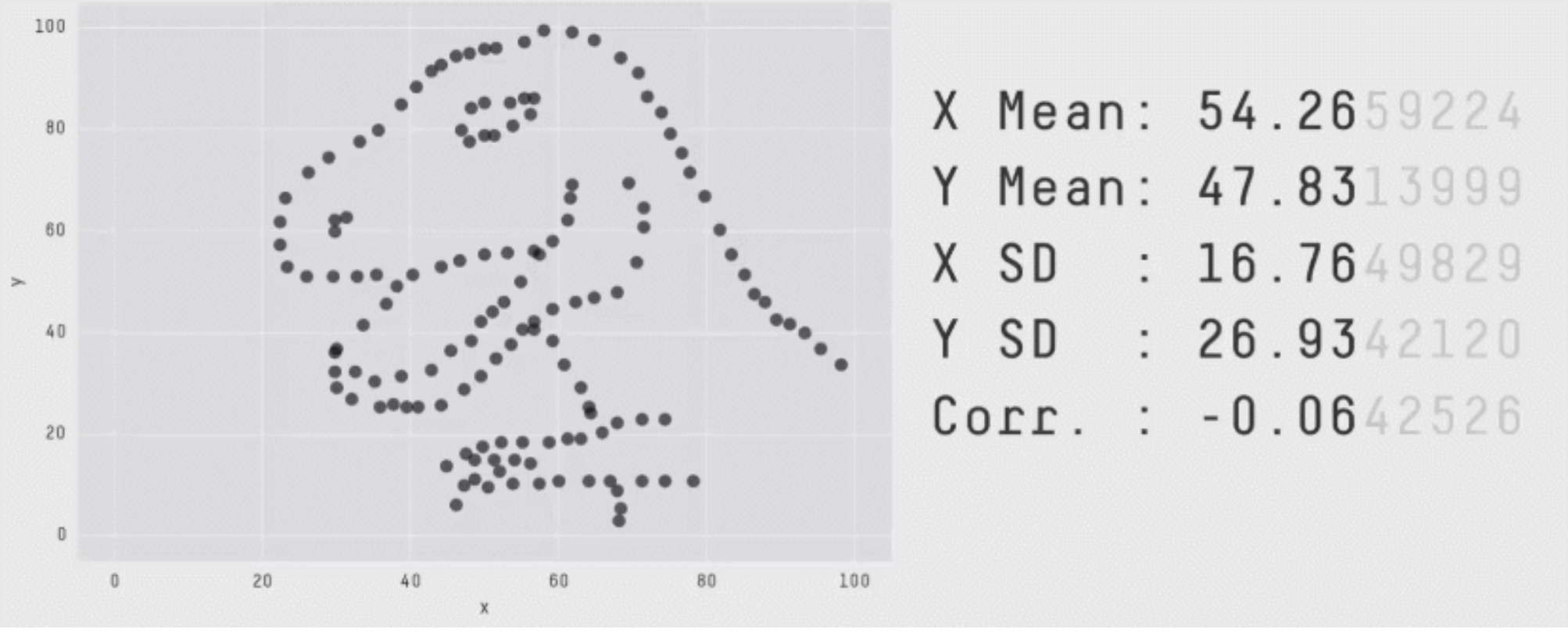
Seems reasonable

```
cor(my_data$x, my_data$y)
```

```
## [1] -0.06447185
```

No correlation

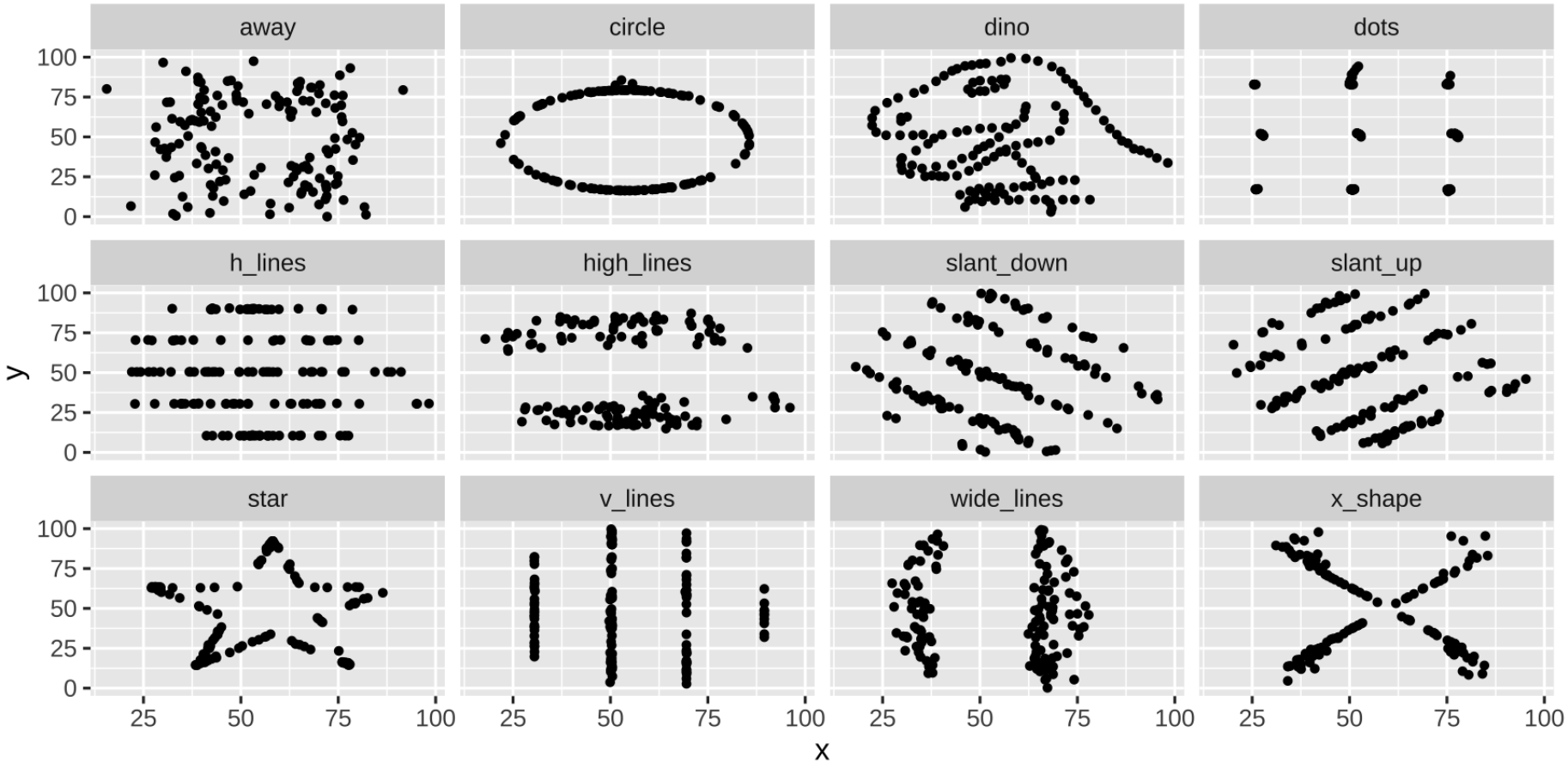
Oh no!



The Datasaurus Dozen

Raw data is not enough

Each of these has the same mean, standard deviation, variance, and correlation



What makes a great visualization?

What makes a great visualization?

Truthful

Functional

Beautiful

Insightful

Enlightening

Alberto Cairo, *The Truthful Art*

What makes a great visualization?

"Graphical excellence is the **well-designed presentation of interesting data**—a matter of substance, of statistics, and of design ... [It] consists of complex ideas communicated with clarity, precision, and efficiency. ... [It] is that which **gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space** ... [It] is nearly always multivariate ... And graphical excellence requires **telling the truth about the data**."

Edward Tufte, *The Visual Display of Quantitative Information*, p. 51

What makes a great visualization?

Good aesthetics

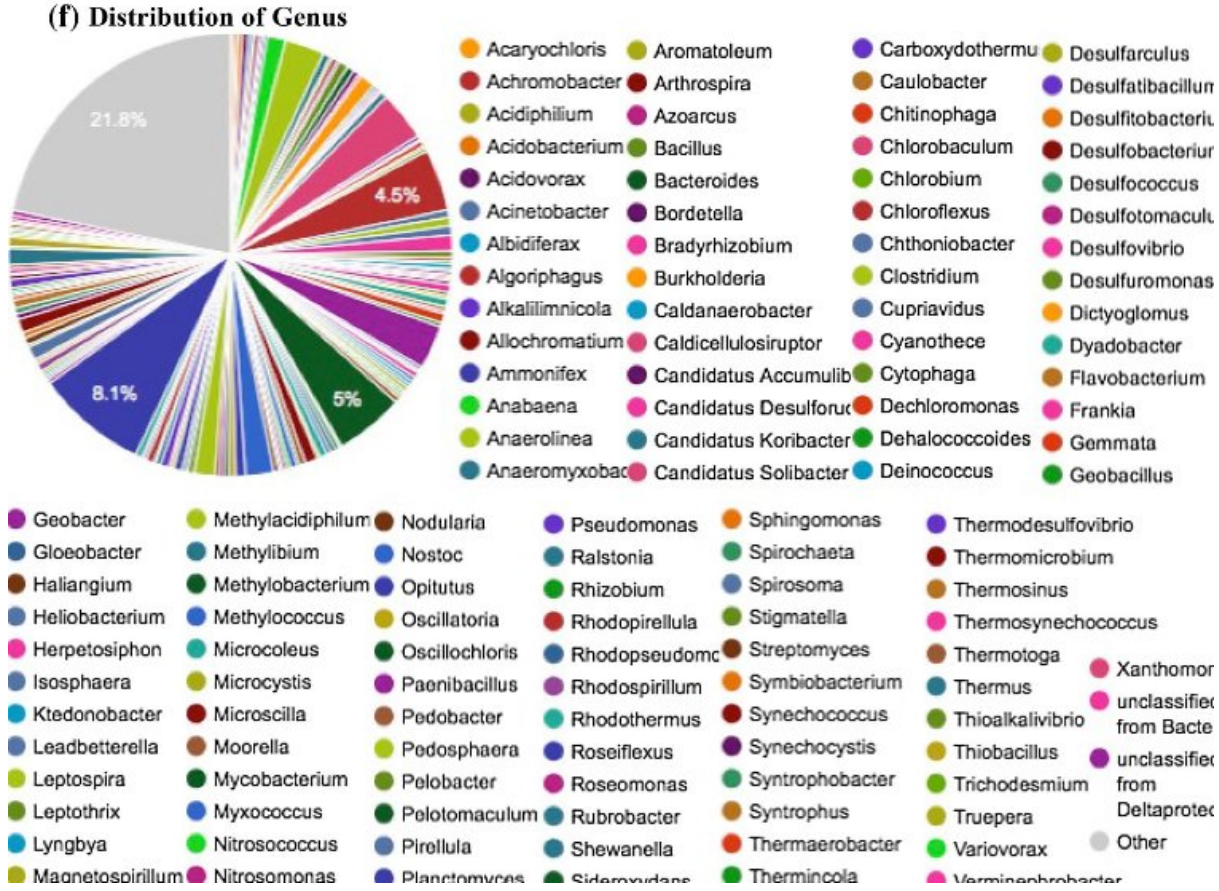
No substantive issues

No perceptual issues

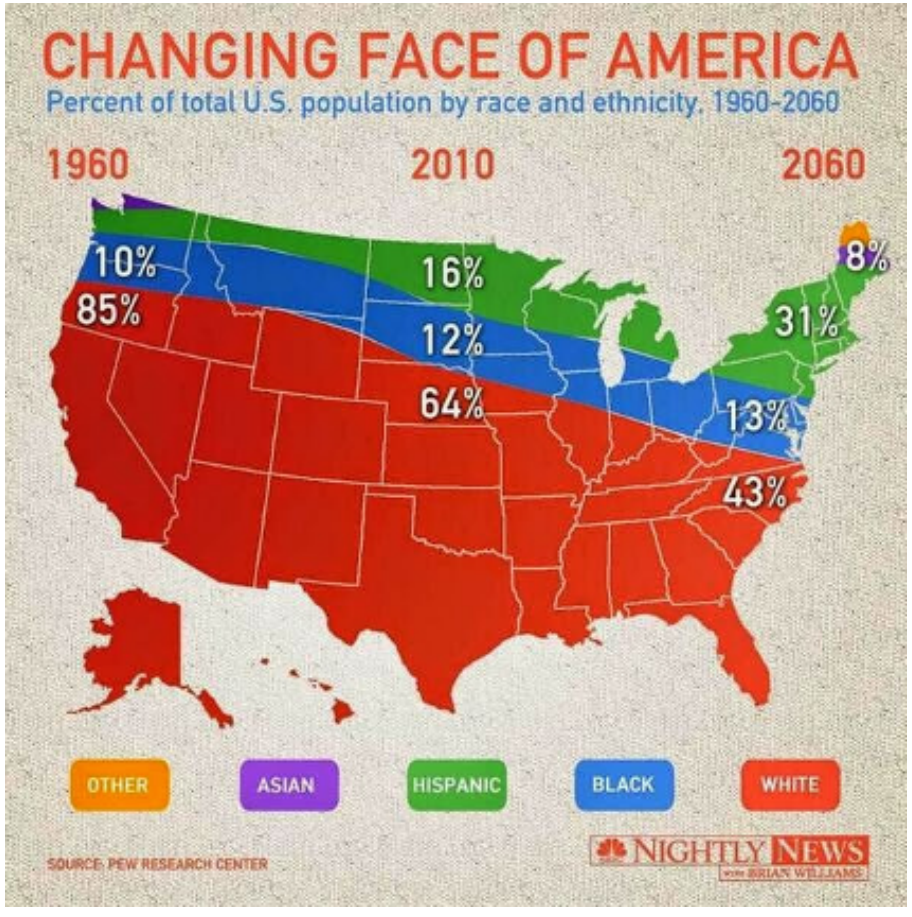
Honesty + good judgment

Kieran Healy, *Data Visualization: A Practical Introduction*

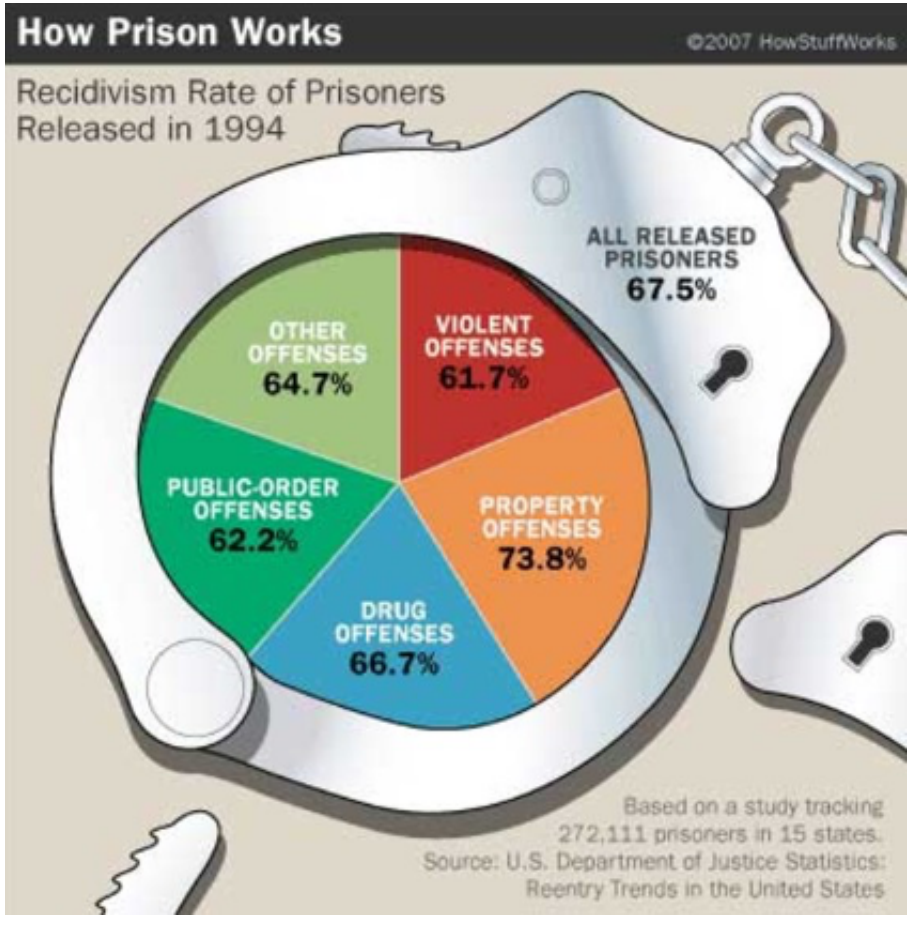
What's wrong?


(f) Distribution of Genus
21.8%
4.5%
8.1%
5%
Acaryochloris
Achromobacter
Acidiphilium
Acidobacterium
Acidovorax
Acinetobacter
Albidiferax
Algoriphagus
Alkalilimnicola
Allochromatium
Ammonifex
Anabaena
Anaerolinea
Anaeromyxobac
Aromatoleum
Arthrospira
Azoarcus
Bacillus
Bacteroides
Bordetella
Bradyrhizobium
Burkholderia
Caldanaerobacter
Caldicellulosiruptor
Candidatus Accumulib
Candidatus Desulforuc
Candidatus Koribacter
Candidatus Solibacter
Carboxydothermu
Caulobacter
Chitinophaga
Chlorobaculum
Chlorobium
Chloroflexus
Chthoniobacter
Clostridium
Cupriavidus
Cyanothece
Cytophaga
Dechloromonas
Dehalococcoides
Deinococcus
Desulfarculus
Desulfatibacillun
Desulfitobacteriu
Desulfobacteriur
Desulfococcus
Desulfotomaculu
Desulfovibrio
Desulfuromonas
Dictyoglomus
Dyadobacter
Flavobacterium
Frankia
Gemmata
Geobacillus
Geobacter
Gloeobacter
Haliangium
Heliobacterium
Herpetosiphon
Isosphaera
Ktedonobacter
Leadbetterella
Leptospira
Leptothrix
Lyngbya
Magnetospirillum
Methylacidiphilum
Methylibium
Methylobacterium
Methylococcus
Microcoleus
Microcystis
Microscilla
Moorella
Mycobacterium
Myxococcus
Nitrosococcus
Nitrosomonas
Nodularia
Nostoc
Opitutus
Oscillatoria
Oscillochloris
Paenibacillus
Pedobacter
Pedosphaera
Pelobacter
Pelotomaculum
Pirellula
Planctomyces
Pseudomonas
Ralstonia
Rhizobium
Rhodopirellula
Rhodopseudomc
Rhodospirillum
Rhodothermus
Roseiflexus
Roseomonas
Rubrobacter
Shewanella
Sideroxydans
Sphingomonas
Spirochaeta
Spirosoma
Stigmatella
Streptomyces
Symbiobacterium
Synechococcus
Synechocystis
Syntrophobacter
Syntrophus
Thermaerobacter
Thermincola
Thermodesulfovibrio
Thermomicrobium
Thermosinus
Thermosynechococcus
Thermotoga
Thermus
Thioalkalivibrio
Thiobacillus
Trichodesmium
Truepera
Variovorax
Xanthomor
unclassifiec
from Bacte
unclassifiec
from
Deltaprotec
Other

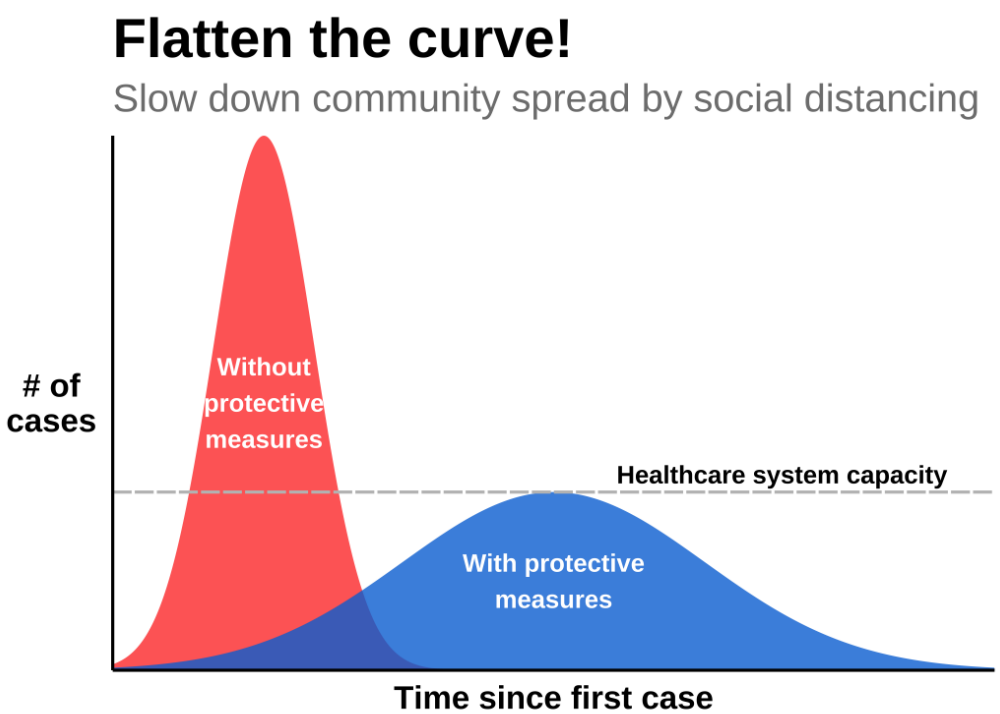
What's wrong?


CHANGING FACE OF AMERICA
Percent of total U.S. population by race and ethnicity, 1960-2060
1960
2010
2060
10%
85%
16%
12%
64%
8%
31%
13%
43%
OTHER
ASIAN
HISPANIC
BLACK
WHITE
SOURCE: PEW RESEARCH CENTER
NIGHTLY NEWS
with BRIAN WILLIAMS

What's wrong?


How Prison Works
©2007 HowStuffWorks
Recidivism Rate of Prisoners Released in 1994
ALL RELEASED PRISONERS
67.5%
OTHER OFFENSES
64.7%
VIOLENT OFFENSES
61.7%
PUBLIC-ORDER OFFENSES
62.2%
PROPERTY OFFENSES
73.8%
DRUG OFFENSES
66.7%
Based on a study tracking 272,111 prisoners in 15 states.
Source: U.S. Department of Justice Statistics: Reentry Trends in the United States

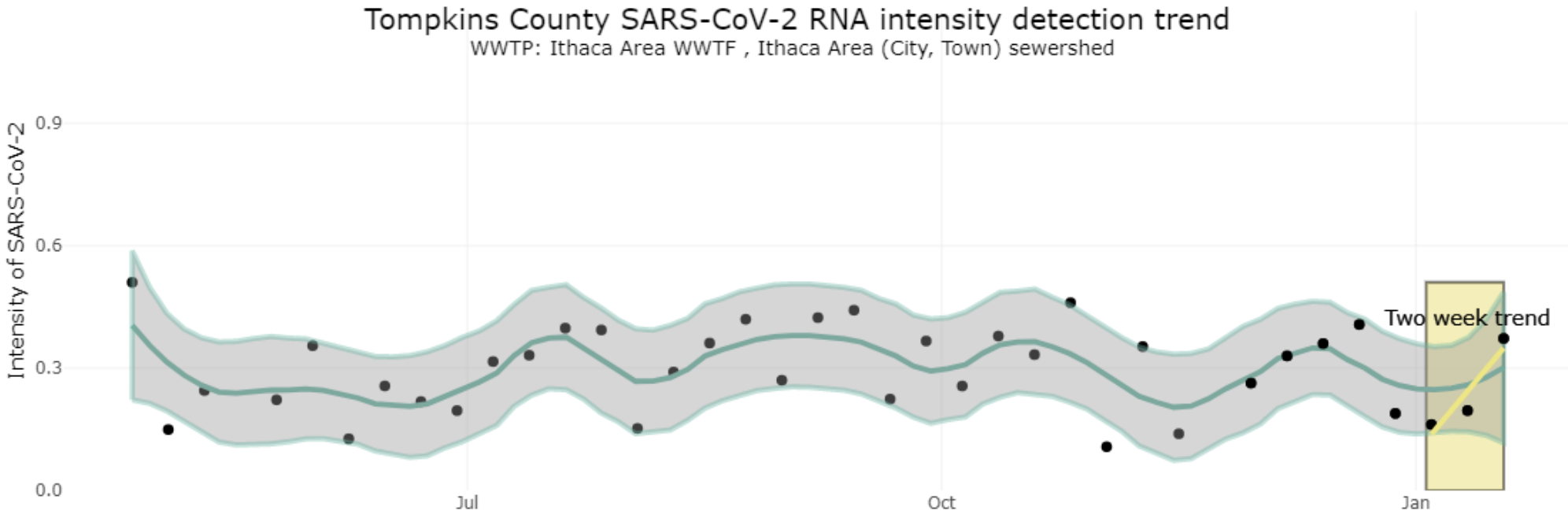
What's right?



Thread by Carl T. Bergstrom

What's wrong? What's right?

Plot of recent COVID levels in Ithaca Area wastewater



Plan for the rest of this week

Office hours:

- Tuesdays 11:00am - 12:00pm: Prof. Gerarden in Warren 466
- Other times by appointment: Prof. Gerarden, at aem2850.youcanbook.me

Thursday:

- Intro to `R`, `RStudio`, and `R Markdown` / `Quarto`
- You will need your computer for coding exercises
- See canvas announcement for instructions to get set up on posit.cloud (formerly rstudio.cloud)

Introductions

Self-introductions

Today we'll do some brief self-introductions

- Goal is to foster a collaborative environment: Who's my professor? Who's in my class?

Self-introductions

1. Name:
    -
2. Where you're from:
    -
3. One other thing about you:
    -

Self-introductions

1. Name:
    - Todd Gerarden
2. Where you're from:
    - Virginia, USA
3. One other thing about you:
    - I rode a bicycle across the country