Advanced Data Journalism: Doing More with R

Module 4: Statistics

Andrew Ba Tran

Know thyself

Understanding the structure of your data opens up the the opportunities for analysis

```
library(tidyverse)
ff <- read_csv("https://github.com/washingtonpost/data-police-shootings/releases/download/v0.1/fatal-police-s

glimpse(ff)
```

```
Rows: 7,666
Columns: 17
$ id                     <dbl> 3, 4, 5, 8, 9, 11, 13, 15, 16, 17, 19, 21, 22,…
$ name                   <chr> "Tim Elliot", "Lewis Lee Lembke", "John Paul Q…
$ date                   <date> 2015-01-02, 2015-01-02, 2015-01-03, 2015-01-0…
$ manner_of_death        <chr> "shot", "shot", "shot and Tasered", "shot", "s…
$ armed                  <chr> "gun", "gun", "unarmed", "toy weapon", "nail g…
$ age                    <dbl> 53, 47, 23, 32, 39, 18, 22, 35, 34, 47, 25, 31…
$ gender                 <chr> "M", "M", "M", "M", "M", "M", "M", "M", "F", "…
$ race                   <chr> "A", "W", "H", "W", "H", "W", "H", "W", "W", "…
$ city                   <chr> "Shelton", "Aloha", "Wichita", "San Francisco"…
$ state                  <chr> "WA", "OR", "KS", "CA", "CO", "OK", "AZ", "KS"…
$ signs_of_mental_illness <lgl> TRUE, FALSE, FALSE, TRUE, FALSE, FALSE, FALSE,…
$ threat_level           <chr> "attack", "attack", "other", "attack", "attack…
$ flee                   <chr> "Not fleeing", "Not fleeing", "Not fleeing", "…
$ body_camera            <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALS…
$ longitude              <dbl> -123.122, -122.892, -97.281, -122.422, -104.69…
$ latitude               <dbl> 47.247, 45.487, 37.695, 37.763, 40.384, 35.877…
$ is_geocoding_exact     <lgl> TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE…
```

Categorical

Frequencies, count

Cross tabs

```
library(lubridate)
```

```r
library(lubridate)

# Frequencies
ff
```

```
# A tibble: 7,666 × 17
      id name    date       manne…¹ armed   age gender race  city  state signs…²
   <dbl> <chr>   <date>     <chr>   <chr> <dbl> <chr>  <chr> <chr> <chr> <lgl>
 1     3 Tim El… 2015-01-02 shot    gun      53 M      A     Shel… WA    TRUE
 2     4 Lewis … 2015-01-02 shot    gun      47 M      W     Aloha OR    FALSE
 3     5 John P… 2015-01-03 shot a… unar…    23 M      H     Wich… KS    FALSE
 4     8 Matthe… 2015-01-04 shot    toy …    32 M      W     San … CA    TRUE
 5     9 Michae… 2015-01-04 shot    nail…    39 M      H     Evans CO    FALSE
 6    11 Kennet… 2015-01-04 shot    gun      18 M      W     Guth… OK    FALSE
 7    13 Kennet… 2015-01-05 shot    gun      22 M      H     Chan… AZ    FALSE
 8    15 Brock … 2015-01-06 shot    gun      35 M      W     Assa… KS    FALSE
 9    16 Autumn… 2015-01-06 shot    unar…    34 F      W     Burl… IA    FALSE
10    17 Leslie… 2015-01-06 shot    toy …    47 M      B     Knox… PA    FALSE
# … with 7,656 more rows, 6 more variables: threat_level <chr>, flee <chr>,
#   body_camera <lgl>, longitude <dbl>, latitude <dbl>,
#   is_geocoding_exact <lgl>, and abbreviated variable names ¹manner_of_death,
#   ²signs_of_mental_illness
# ℹ Use `print(n = ...)` to see more rows, and `colnames()` to see all variable names
```

```r
library(lubridate)

# Frequencies
ff %>%
  group_by(signs_of_mental_illness)
```

```
# A tibble: 7,666 × 17
# Groups:   signs_of_mental_illness [2]
      id name    date       manne…¹ armed   age gender race  city  state signs…²
   <dbl> <chr>   <date>     <chr>   <chr> <dbl> <chr>  <chr> <chr> <chr> <lgl>
 1     3 Tim El… 2015-01-02 shot    gun      53 M      A     Shel… WA    TRUE
 2     4 Lewis … 2015-01-02 shot    gun      47 M      W     Aloha OR    FALSE
 3     5 John P… 2015-01-03 shot a… unar…    23 M      H     Wich… KS    FALSE
 4     8 Matthe… 2015-01-04 shot    toy …    32 M      W     San … CA    TRUE
 5     9 Michae… 2015-01-04 shot    nail…    39 M      H     Evans CO    FALSE
 6    11 Kennet… 2015-01-04 shot    gun      18 M      W     Guth… OK    FALSE
 7    13 Kennet… 2015-01-05 shot    gun      22 M      H     Chan… AZ    FALSE
 8    15 Brock … 2015-01-06 shot    gun      35 M      W     Assa… KS    FALSE
 9    16 Autumn… 2015-01-06 shot    unar…    34 F      W     Burl… IA    FALSE
10    17 Leslie… 2015-01-06 shot    toy …    47 M      B     Knox… PA    FALSE
# … with 7,656 more rows, 6 more variables: threat_level <chr>, flee <chr>,
#   body_camera <lgl>, longitude <dbl>, latitude <dbl>,
#   is_geocoding_exact <lgl>, and abbreviated variable names ¹manner_of_death,
#   ²signs_of_mental_illness
# ℹ Use `print(n = ...)` to see more rows, and `colnames()` to see all variable names
```

```r
library(lubridate)

# Frequencies
ff %>%
  group_by(signs_of_mental_illness) %>%
  summarize(total=n())
```

```
# A tibble: 2 × 2
  signs_of_mental_illness total
  <lgl>                   <int>
1 FALSE                    6042
2 TRUE                     1624
```

```r
library(lubridate)

# Frequencies
ff %>%
  group_by(signs_of_mental_illness) %>%
  summarize(total=n())

# Cross tabs
ff
```

```
# A tibble: 2 × 2
  signs_of_mental_illness total
  <lgl>                   <int>
1 FALSE                    6042
2 TRUE                     1624

# A tibble: 7,666 × 17
      id name    date       manne…¹ armed   age gender race  city  state signs…²
   <dbl> <chr>   <date>     <chr>   <chr> <dbl> <chr>  <chr> <chr> <chr> <lgl>
 1     3 Tim El… 2015-01-02 shot    gun      53 M      A     Shel… WA    TRUE
 2     4 Lewis … 2015-01-02 shot    gun      47 M      W     Aloha OR    FALSE
 3     5 John P… 2015-01-03 shot a… unar…    23 M      H     Wich… KS    FALSE
 4     8 Matthe… 2015-01-04 shot    toy …    32 M      W     San … CA    TRUE
 5     9 Michae… 2015-01-04 shot    nail…    39 M      H     Evans CO    FALSE
 6    11 Kennet… 2015-01-04 shot    gun      18 M      W     Guth… OK    FALSE
 7    13 Kennet… 2015-01-05 shot    gun      22 M      H     Chan… AZ    FALSE
 8    15 Brock … 2015-01-06 shot    gun      35 M      W     Assa… KS    FALSE
 9    16 Autumn… 2015-01-06 shot    unar…    34 F      W     Burl… IA    FALSE
10    17 Leslie… 2015-01-06 shot    toy …    47 M      B     Knox… PA    FALSE
# … with 7,656 more rows, 6 more variables: threat_level <chr>, flee <chr>,
#   body_camera <lgl>, longitude <dbl>, latitude <dbl>,
#   is_geocoding_exact <lgl>, and abbreviated variable names ¹manner_of_death,
#   ²signs_of_mental_illness
# ℹ Use `print(n = ...)` to see more rows, and `colnames()` to see all variable names
```

```r
library(lubridate)

# Frequencies
ff %>%
  group_by(signs_of_mental_illness) %>%
  summarize(total=n())

# Cross tabs
ff %>%
  group_by(signs_of_mental_illness, armed)
```

```
# A tibble: 2 × 2
  signs_of_mental_illness total
  <lgl>                   <int>
1 FALSE                    6042
2 TRUE                     1624

# A tibble: 7,666 × 17
# Groups:   signs_of_mental_illness, armed [157]
      id name    date       manne…¹ armed   age gender race  city  state signs…²
   <dbl> <chr>   <date>     <chr>   <chr> <dbl> <chr>  <chr> <chr> <chr> <lgl>
 1     3 Tim El… 2015-01-02 shot    gun      53 M      A     Shel… WA    TRUE
 2     4 Lewis … 2015-01-02 shot    gun      47 M      W     Aloha OR    FALSE
 3     5 John P… 2015-01-03 shot a… unar…    23 M      H     Wich… KS    FALSE
 4     8 Matthe… 2015-01-04 shot    toy …    32 M      W     San … CA    TRUE
 5     9 Michae… 2015-01-04 shot    nail…    39 M      H     Evans CO    FALSE
 6    11 Kennet… 2015-01-04 shot    gun      18 M      W     Guth… OK    FALSE
 7    13 Kennet… 2015-01-05 shot    gun      22 M      H     Chan… AZ    FALSE
 8    15 Brock … 2015-01-06 shot    gun      35 M      W     Assa… KS    FALSE
 9    16 Autumn… 2015-01-06 shot    unar…    34 F      W     Burl… IA    FALSE
10    17 Leslie… 2015-01-06 shot    toy …    47 M      B     Knox… PA    FALSE
# … with 7,656 more rows, 6 more variables: threat_level <chr>, flee <chr>,
#   body_camera <lgl>, longitude <dbl>, latitude <dbl>,
#   is_geocoding_exact <lgl>, and abbreviated variable names ¹manner_of_death,
#   ²signs_of_mental_illness
# ℹ Use `print(n = ...)` to see more rows, and `colnames()` to see all variable names
```

```r
library(lubridate)

# Frequencies
ff %>%
  group_by(signs_of_mental_illness) %>%
  summarize(total=n())

# Cross tabs
ff %>%
  group_by(signs_of_mental_illness, armed) %>%
  summarize(total=n())
```

```
# A tibble: 2 × 2
  signs_of_mental_illness total
  <lgl>                   <int>
1 FALSE                    6042
2 TRUE                     1624

# A tibble: 157 × 3
# Groups:   signs_of_mental_illness [2]
   signs_of_mental_illness armed                          total
   <lgl>                   <chr>                          <int>
 1 FALSE                   air conditioner                    1
 2 FALSE                   air pistol                         2
 3 FALSE                   Airsoft pistol                     5
 4 FALSE                   ax                                14
 5 FALSE                   ax and machete                     1
 6 FALSE                   baseball bat                      12
 7 FALSE                   baseball bat and fireplace poker   1
 8 FALSE                   baton                              4
 9 FALSE                   BB gun                             9
10 FALSE                   BB gun and vehicle                 1
# … with 147 more rows
# ℹ Use `print(n = ...)` to see more rows
```

Continuous data

- **Mean**
- **Median**
- **Range**
- **Rank**

```
ff
```

```
# A tibble: 7,666 × 17
      id name    date       manne…¹ armed   age gender race  city  state signs…²
   <dbl> <chr>   <date>     <chr>   <chr> <dbl> <chr>  <chr> <chr> <chr> <lgl>
 1     3 Tim El… 2015-01-02 shot    gun      53 M      A     Shel… WA    TRUE
 2     4 Lewis … 2015-01-02 shot    gun      47 M      W     Aloha OR    FALSE
 3     5 John P… 2015-01-03 shot a… unar…    23 M      H     Wich… KS    FALSE
 4     8 Matthe… 2015-01-04 shot    toy …    32 M      W     San … CA    TRUE
 5     9 Michae… 2015-01-04 shot    nail…    39 M      H     Evans CO    FALSE
 6    11 Kennet… 2015-01-04 shot    gun      18 M      W     Guth… OK    FALSE
 7    13 Kennet… 2015-01-05 shot    gun      22 M      H     Chan… AZ    FALSE
 8    15 Brock … 2015-01-06 shot    gun      35 M      W     Assa… KS    FALSE
 9    16 Autumn… 2015-01-06 shot    unar…    34 F      W     Burl… IA    FALSE
10    17 Leslie… 2015-01-06 shot    toy …    47 M      B     Knox… PA    FALSE
# … with 7,656 more rows, 6 more variables: threat_level <chr>, flee <chr>,
#   body_camera <lgl>, longitude <dbl>, latitude <dbl>,
#   is_geocoding_exact <lgl>, and abbreviated variable names ¹manner_of_death,
#   ²signs_of_mental_illness
# ℹ Use `print(n = ...)` to see more rows, and `colnames()` to see all variable names
```

```
ff %>%
  summarize(mean=mean(age, na.rm=T),
            median=median(age, na.rm=T),
            min_age=min(age, na.rm=T),
            max_age=max(age, na.rm=T))
```

```
# A tibble: 1 × 4
   mean median min_age max_age
  <dbl>  <dbl>   <dbl>   <dbl>
1  37.2     35       2      92
```

Continuous data (MORE!)

- **N-tiles**
- **Rates**
- **Correlation**
- **Regression**

```
state_pop <- read_csv("https://docs.google.com/sprea
```

```
state_pop <- read_csv("https://docs.google.com/sprea

glimpse(state_pop)
```

```
Rows: 51
Columns: 3
$ statefull  <chr> "Alaska", "Alabama", "Arkansas", "Arizona", "California", "…
$ state      <chr> "AK", "AL", "AR", "AZ", "CA", "CO", "CT", "DC", "DE", "FL",…
$ population <dbl> 736732, 4849377, 2966369, 6731484, 38802500, 5355866, 35966…
```

```
state_pop <- read_csv("https://docs.google.com/sprea

glimpse(state_pop)

ff
```

```
Rows: 51
Columns: 3
$ statefull  <chr> "Alaska", "Alabama", "Arkansas", "Arizona", "California", "…
$ state      <chr> "AK", "AL", "AR", "AZ", "CA", "CO", "CT", "DC", "DE", "FL",…
$ population <dbl> 736732, 4849377, 2966369, 6731484, 38802500, 5355866, 35966…

# A tibble: 7,666 × 17
      id name    date       manne…¹ armed   age gender race  city  state signs…²
   <dbl> <chr>   <date>     <chr>   <chr> <dbl> <chr>  <chr> <chr> <chr> <lgl>
 1     3 Tim El… 2015-01-02 shot    gun      53 M      A     Shel… WA    TRUE
 2     4 Lewis … 2015-01-02 shot    gun      47 M      W     Aloha OR    FALSE
 3     5 John P… 2015-01-03 shot a… unar…    23 M      H     Wich… KS    FALSE
 4     8 Matthe… 2015-01-04 shot    toy …    32 M      W     San … CA    TRUE
 5     9 Michae… 2015-01-04 shot    nail…    39 M      H     Evans CO    FALSE
 6    11 Kennet… 2015-01-04 shot    gun      18 M      W     Guth… OK    FALSE
 7    13 Kennet… 2015-01-05 shot    gun      22 M      H     Chan… AZ    FALSE
 8    15 Brock … 2015-01-06 shot    gun      35 M      W     Assa… KS    FALSE
 9    16 Autumn… 2015-01-06 shot    unar…    34 F      W     Burl… IA    FALSE
10    17 Leslie… 2015-01-06 shot    toy …    47 M      B     Knox… PA    FALSE
# … with 7,656 more rows, 6 more variables: threat_level <chr>, flee <chr>,
#   body_camera <lgl>, longitude <dbl>, latitude <dbl>,
#   is_geocoding_exact <lgl>, and abbreviated variable names ¹manner_of_death,
#   ²signs_of_mental_illness
# ℹ Use `print(n = ...)` to see more rows, and `colnames()` to see all variable names
```

```
state_pop <- read_csv("https://docs.google.com/sprea

glimpse(state_pop)


ff %>%
  group_by(state)
```

```
Rows: 51
Columns: 3
$ statefull  <chr> "Alaska", "Alabama", "Arkansas", "Arizona", "California", "…
$ state      <chr> "AK", "AL", "AR", "AZ", "CA", "CO", "CT", "DC", "DE", "FL",…
$ population <dbl> 736732, 4849377, 2966369, 6731484, 38802500, 5355866, 35966…

# A tibble: 7,666 × 17
# Groups:   state [51]
      id name    date       manne…¹ armed   age gender race  city  state signs…²
   <dbl> <chr>   <date>     <chr>   <chr> <dbl> <chr>  <chr> <chr> <chr> <lgl>
 1     3 Tim El… 2015-01-02 shot    gun      53 M      A     Shel… WA    TRUE
 2     4 Lewis … 2015-01-02 shot    gun      47 M      W     Aloha OR    FALSE
 3     5 John P… 2015-01-03 shot a… unar…    23 M      H     Wich… KS    FALSE
 4     8 Matthe… 2015-01-04 shot    toy …    32 M      W     San … CA    TRUE
 5     9 Michae… 2015-01-04 shot    nail…    39 M      H     Evans CO    FALSE
 6    11 Kennet… 2015-01-04 shot    gun      18 M      W     Guth… OK    FALSE
 7    13 Kennet… 2015-01-05 shot    gun      22 M      H     Chan… AZ    FALSE
 8    15 Brock … 2015-01-06 shot    gun      35 M      W     Assa… KS    FALSE
 9    16 Autumn… 2015-01-06 shot    unar…    34 F      W     Burl… IA    FALSE
10    17 Leslie… 2015-01-06 shot    toy …    47 M      B     Knox… PA    FALSE
# … with 7,656 more rows, 6 more variables: threat_level <chr>, flee <chr>,
#   body_camera <lgl>, longitude <dbl>, latitude <dbl>,
#   is_geocoding_exact <lgl>, and abbreviated variable names ¹manner_of_death,
#   ²signs_of_mental_illness
# ℹ Use `print(n = ...)` to see more rows, and `colnames()` to see all variable names
```

```
state_pop <- read_csv("https://docs.google.com/sprea

glimpse(state_pop)


ff %>%
  group_by(state) %>%
  summarize(shootings=n())
```

```
Rows: 51
Columns: 3
$ statefull  <chr> "Alaska", "Alabama", "Arkansas", "Arizona", "California", "…
$ state      <chr> "AK", "AL", "AR", "AZ", "CA", "CO", "CT", "DC", "DE", "FL",…
$ population <dbl> 736732, 4849377, 2966369, 6731484, 38802500, 5355866, 35966…

# A tibble: 51 × 2
   state shootings
   <chr>     <int>
 1 AK           52
 2 AL          144
 3 AR          109
 4 AZ          348
 5 CA         1109
 6 CO          278
 7 CT           22
 8 DC           24
 9 DE           17
10 FL          492
# … with 41 more rows
# ℹ Use `print(n = ...)` to see more rows
```

```
state_pop <- read_csv("https://docs.google.com/sprea

glimpse(state_pop)


ff %>%
  group_by(state) %>%
  summarize(shootings=n()) %>%
  left_join(state_pop)
```

```
Rows: 51
Columns: 3
$ statefull  <chr> "Alaska", "Alabama", "Arkansas", "Arizona", "California", "…
$ state      <chr> "AK", "AL", "AR", "AZ", "CA", "CO", "CT", "DC", "DE", "FL",…
$ population <dbl> 736732, 4849377, 2966369, 6731484, 38802500, 5355866, 35966…

# A tibble: 51 × 4
   state shootings statefull            population
   <chr>     <int> <chr>                     <dbl>
 1 AK           52 Alaska                   736732
 2 AL          144 Alabama                 4849377
 3 AR          109 Arkansas                2966369
 4 AZ          348 Arizona                 6731484
 5 CA         1109 California             38802500
 6 CO          278 Colorado                5355866
 7 CT           22 Connecticut             3596677
 8 DC           24 District of Columbia     658893
 9 DE           17 Delaware                 935614
10 FL          492 Florida                19893297
# … with 41 more rows
# ℹ Use `print(n = ...)` to see more rows
```

```
state_pop <- read_csv("https://docs.google.com/sprea

glimpse(state_pop)


ff %>%
  group_by(state) %>%
  summarize(shootings=n()) %>%
  left_join(state_pop) %>%
  mutate(per_100k=shootings/population*100000)
```

```
Rows: 51
Columns: 3
$ statefull  <chr> "Alaska", "Alabama", "Arkansas", "Arizona", "California", "…
$ state      <chr> "AK", "AL", "AR", "AZ", "CA", "CO", "CT", "DC", "DE", "FL",…
$ population <dbl> 736732, 4849377, 2966369, 6731484, 38802500, 5355866, 35966…

# A tibble: 51 × 5
   state shootings statefull            population per_100k
   <chr>     <int> <chr>                     <dbl>    <dbl>
 1 AK           52 Alaska                   736732    7.06
 2 AL          144 Alabama                 4849377    2.97
 3 AR          109 Arkansas                2966369    3.67
 4 AZ          348 Arizona                 6731484    5.17
 5 CA         1109 California             38802500    2.86
 6 CO          278 Colorado                5355866    5.19
 7 CT           22 Connecticut             3596677    0.612
 8 DC           24 District of Columbia     658893    3.64
 9 DE           17 Delaware                 935614    1.82
10 FL          492 Florida                19893297    2.47
# … with 41 more rows
# ℹ Use `print(n = ...)` to see more rows
```

```r
state_pop <- read_csv("https://docs.google.com/sprea

glimpse(state_pop)


ff %>%
  group_by(state) %>%
  summarize(shootings=n()) %>%
  left_join(state_pop) %>%
  mutate(per_100k=shootings/population*100000) %>%
  mutate(ntile=ntile(population, 4))
```

```
Rows: 51
Columns: 3
$ statefull  <chr> "Alaska", "Alabama", "Arkansas", "Arizona", "California", "…
$ state      <chr> "AK", "AL", "AR", "AZ", "CA", "CO", "CT", "DC", "DE", "FL",…
$ population <dbl> 736732, 4849377, 2966369, 6731484, 38802500, 5355866, 35966…

# A tibble: 51 × 6
   state shootings statefull            population per_100k ntile
   <chr>     <int> <chr>                     <dbl>    <dbl> <int>
 1 AK           52 Alaska                   736732    7.06      1
 2 AL          144 Alabama                 4849377    2.97      3
 3 AR          109 Arkansas                2966369    3.67      2
 4 AZ          348 Arizona                 6731484    5.17      3
 5 CA         1109 California             38802500    2.86      4
 6 CO          278 Colorado                5355866    5.19      3
 7 CT           22 Connecticut             3596677    0.612     2
 8 DC           24 District of Columbia     658893    3.64      1
 9 DE           17 Delaware                 935614    1.82      1
10 FL          492 Florida                19893297    2.47      4
# … with 41 more rows
# ℹ Use `print(n = ...)` to see more rows
```

```r
state_pop <- read_csv("https://docs.google.com/sprea

glimpse(state_pop)


ff %>%
  group_by(state) %>%
  summarize(shootings=n()) %>%
  left_join(state_pop) %>%
  mutate(per_100k=shootings/population*100000) %>%
  mutate(ntile=ntile(population, 4)) %>%
  group_by(ntile)
```

```
Rows: 51
Columns: 3
$ statefull  <chr> "Alaska", "Alabama", "Arkansas", "Arizona", "California", "…
$ state      <chr> "AK", "AL", "AR", "AZ", "CA", "CO", "CT", "DC", "DE", "FL",…
$ population <dbl> 736732, 4849377, 2966369, 6731484, 38802500, 5355866, 35966…

# A tibble: 51 × 6
# Groups:   ntile [4]
   state shootings statefull            population per_100k ntile
   <chr>     <int> <chr>                     <dbl>    <dbl> <int>
 1 AK           52 Alaska                   736732    7.06      1
 2 AL          144 Alabama                 4849377    2.97      3
 3 AR          109 Arkansas                2966369    3.67      2
 4 AZ          348 Arizona                 6731484    5.17      3
 5 CA         1109 California             38802500    2.86      4
 6 CO          278 Colorado                5355866    5.19      3
 7 CT           22 Connecticut             3596677    0.612     2
 8 DC           24 District of Columbia     658893    3.64      1
 9 DE           17 Delaware                 935614    1.82      1
10 FL          492 Florida                19893297    2.47      4
# … with 41 more rows
# ℹ Use `print(n = ...)` to see more rows
```

```r
state_pop <- read_csv("https://docs.google.com/sprea

glimpse(state_pop)


ff %>%
  group_by(state) %>%
  summarize(shootings=n()) %>%
  left_join(state_pop) %>%
  mutate(per_100k=shootings/population*100000) %>%
  mutate(ntile=ntile(population, 4)) %>%
  group_by(ntile) %>%
  summarize(mean=mean(per_100k))
```

```
Rows: 51
Columns: 3
$ statefull  <chr> "Alaska", "Alabama", "Arkansas", "Arizona", "California", "…
$ state      <chr> "AK", "AL", "AR", "AZ", "CA", "CO", "CT", "DC", "DE", "FL",…
$ population <dbl> 736732, 4849377, 2966369, 6731484, 38802500, 5355866, 35966…

# A tibble: 4 × 2
  ntile  mean
  <int> <dbl>
1     1  3.07
2     2  3.41
3     3  2.83
4     4  1.82
```

Exploring relationships between variables

Linear/Logistic regression

Measure of relationship between variables

Useful for inference (relationship) and prediction

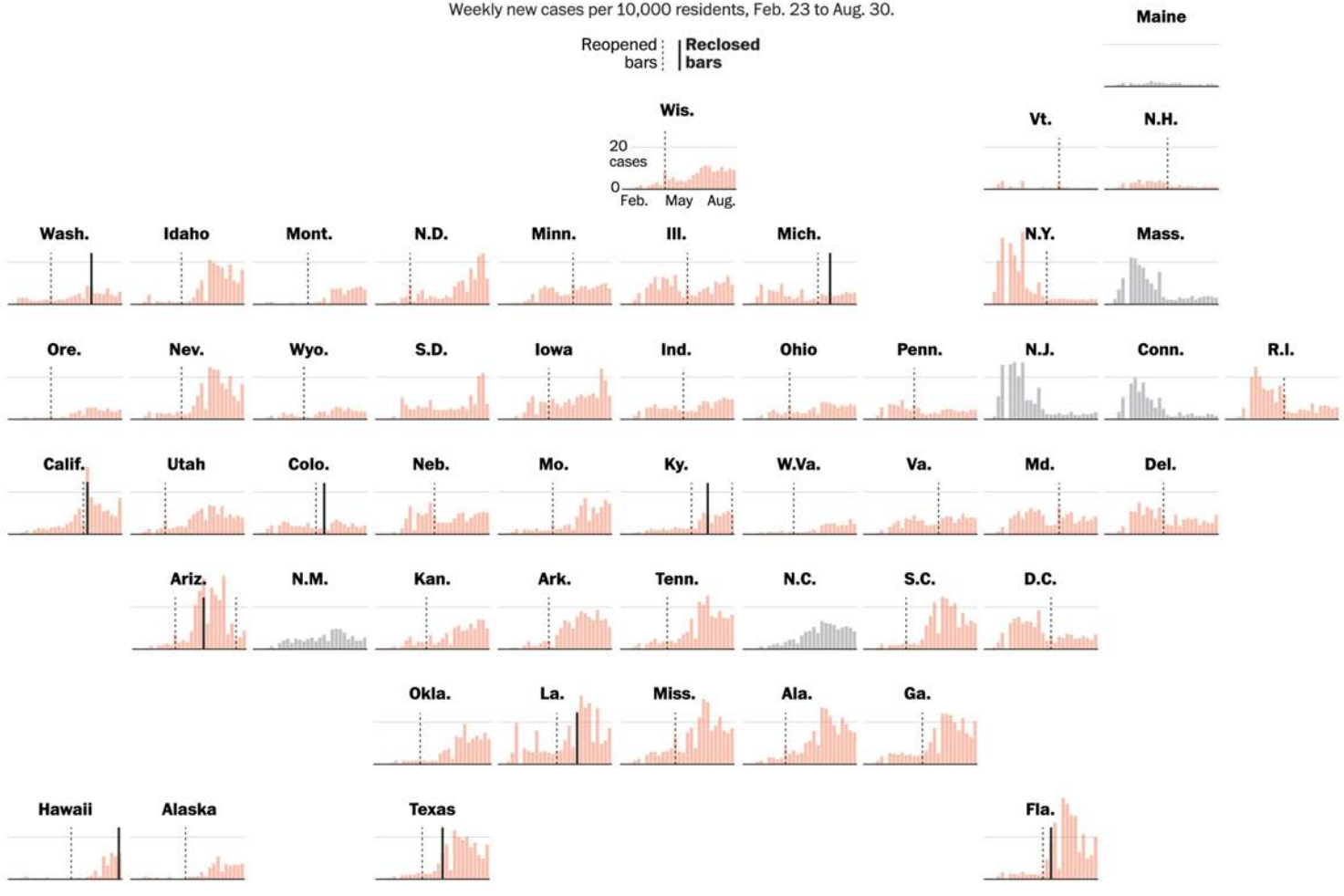
- **Inference** asks: how accurate is our estimate of the relationship between variables
- **Prediction** asks: how accurately can be predict the outcome variable

Linear when continuous, logistic when discrete, categorical, or binary

Correlation

Coronavirus cases reported in the weeks before and after states reopened bars

Weekly new cases per 10,000 residents, Feb. 23 to Aug. 30.



Note: As of Sept. 14, Connecticut, Maine, Massachusetts, North Carolina, New Jersey and New Mexico have not reopened bars. South Dakota has no statewide restrictions. In some states, restrictions vary by region.


Bar foot traffic change after reopening vs. lagged change in covid cases
Change in foot traffic 1 week after bars reopened
Change in cases per capita 3-4 weeks after bars reopened
75
50
25
0
0
100
200
NV
AK
ND
MS
AL
WY
AZ
WI
MO
TN
OK
TX
ID
RI
OH
MI
GA
OR
DE
NE
WV
FL
SC
CA
PA
IA
NH
MN
VA
WA
VT
MT
LA
UT
IN
DC
IL
HI
CO
KS
MD
AR

One decision appears to be riskier than the other, according to an analysis of cellphone and coronavirus case data by The Washington Post.

States that have reopened bars experienced a doubling in the rate of coronavirus cases three weeks after the opening of doors, on average. The Post analysis — using data provided by SafeGraph, a company that aggregates cellphone location information — found a statistically significant national relationship between foot traffic to bars one week after they reopened and an increase in cases three weeks later.

The analysis of the cellphone data suggests there is not as strong a relationship between the reopening of restaurants and a rise in cases, nor with bar foot traffic and cases over time, except for a handful of states.

Correlation != Causation

Linear modeling

INVISIBLE

Countries' climate pledges built on flawed data, Post investigation finds

3. Creating a model to estimate what emissions each country would have reported in 2019, if they only reported in an earlier year

- To overcome missing data, The Post used a linear regression technique to model what countries would have reported in 2019, measuring past years of reports against independent estimates from Minx et al., a research effort that has totaled each country's greenhouse gases.

With great power comes great responsibility

Most of the time, the uncomplicated process is better for the reader to grasp

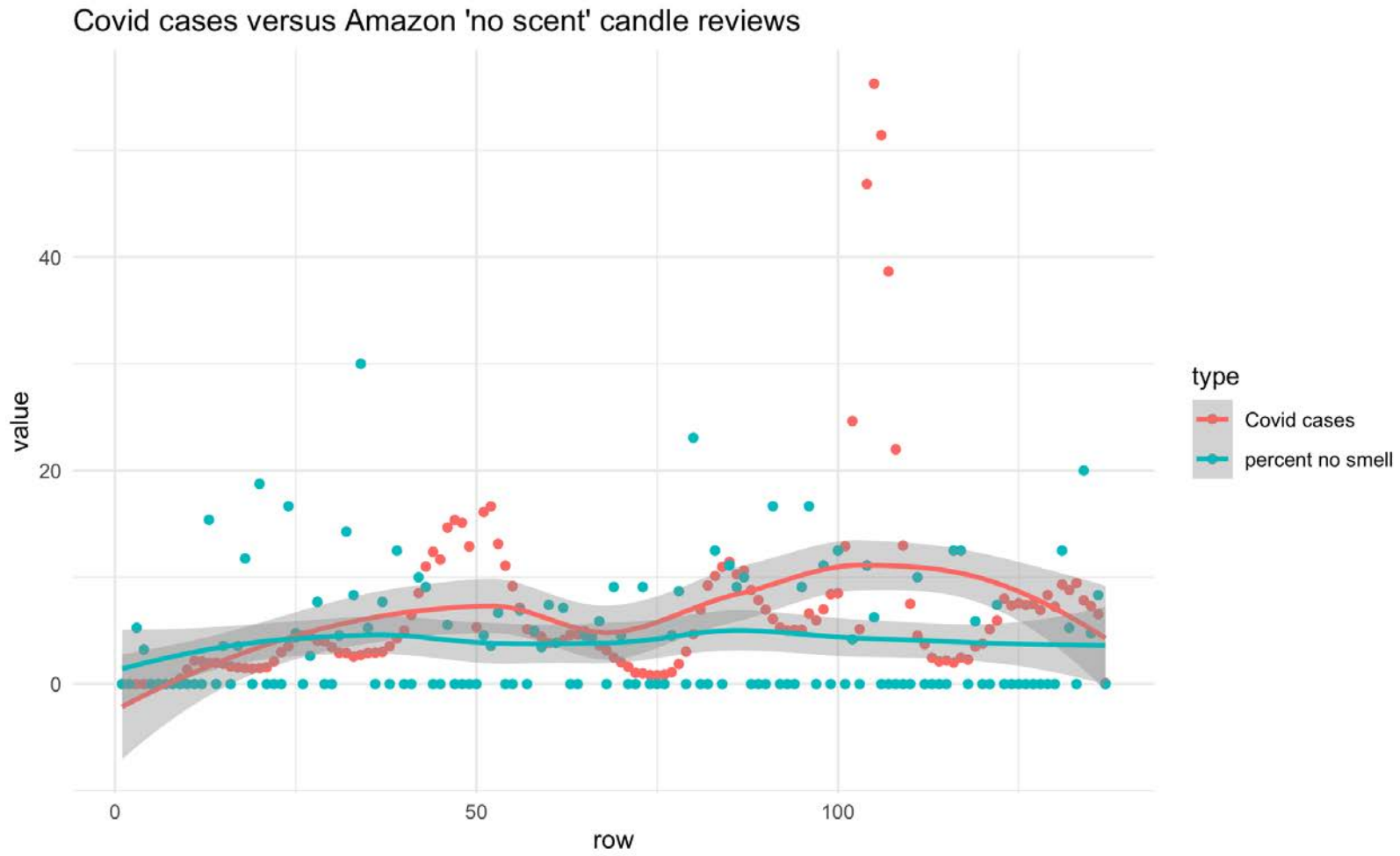
Explain your methodology

Save the complicated numbers for the graphics

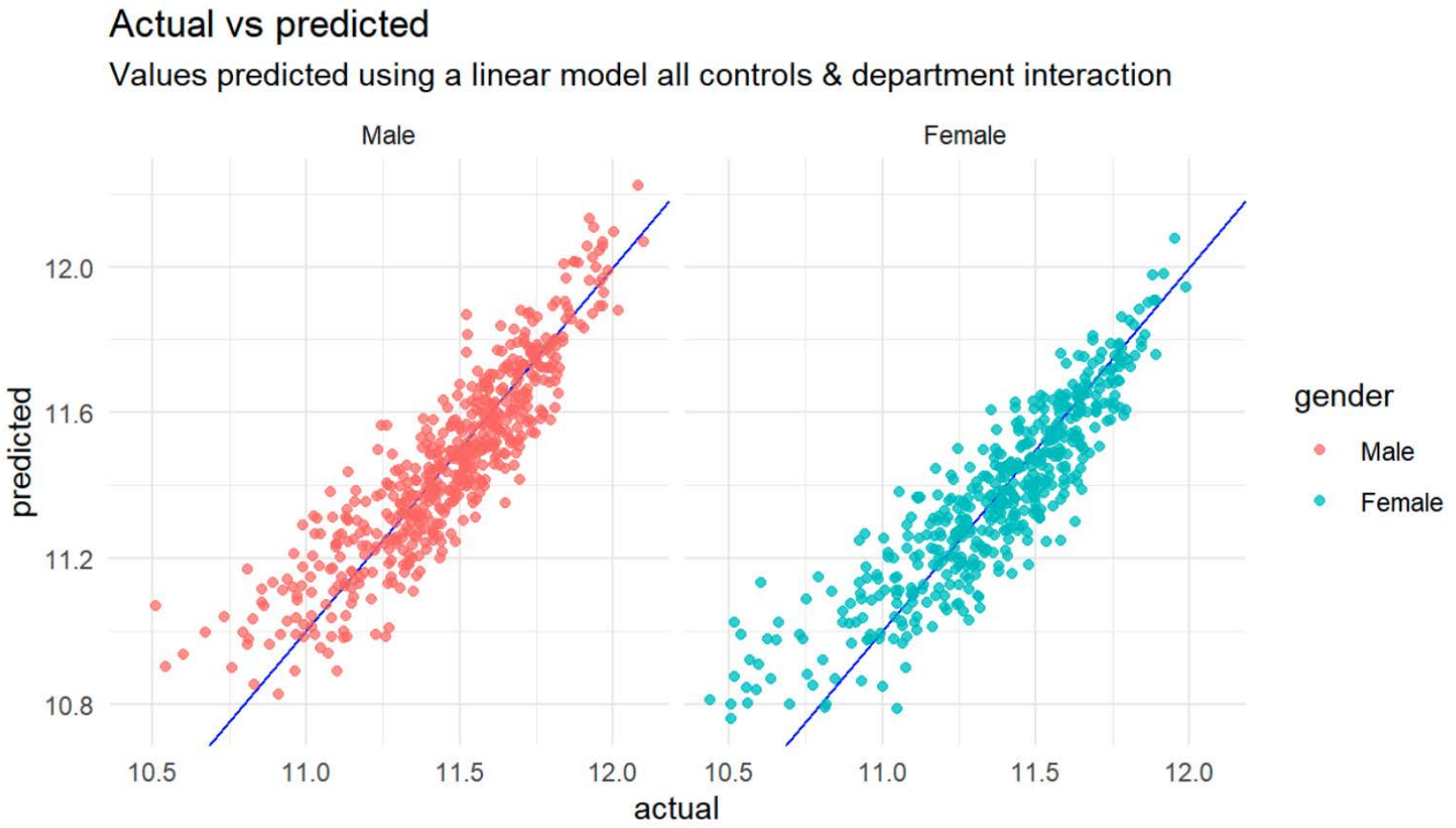
Run your process by experts

Avoid spurious correlations

Explaining statistical significance is difficult


Covid cases versus Amazon 'no scent' candle reviews
value
40
20
0
0
50
100
row
type
Covid cases
percent no smell

jobTitle	gender	age	perfEval	edu	dept	seniority	basePay	bonus
Graphic Designer	Female	18	5	College	Operations	2	42363	9938
Software Engineer	Male	21	5	College	Management	5	108476	11128
Warehouse Associate	Female	19	4	PhD	Administration	5	90208	9268
Software Engineer	Male	20	5	Masters	Sales	4	108080	10154
Graphic Designer	Male	26	5	Masters	Engineering	5	99464	9319
IT	Female	20	5	PhD	Operations	4	70890	10126
Graphic Designer	Female	20	5	College	Sales	4	67585	10541
Software Engineer	Male	18	4	PhD	Engineering	5	97523	10240
Graphic Designer	Female	33	5	High School	Engineering	5	112976	9836
Sales Associate	Female	35	5	College	Engineering	5	106524	9941
Graphic Designer	Male	24	5	PhD	Engineering	5	102261	10212
Driver	Female	18	5	College	Management	3	62759	10124
Financial Analyst	Female	19	5	College	Sales	3	84007	8990
Warehouse Associate	Female	30	5	Masters	Administration	5	86220	9583
Warehouse Associate	Female	35	5	PhD	Operations	4	95584	9745

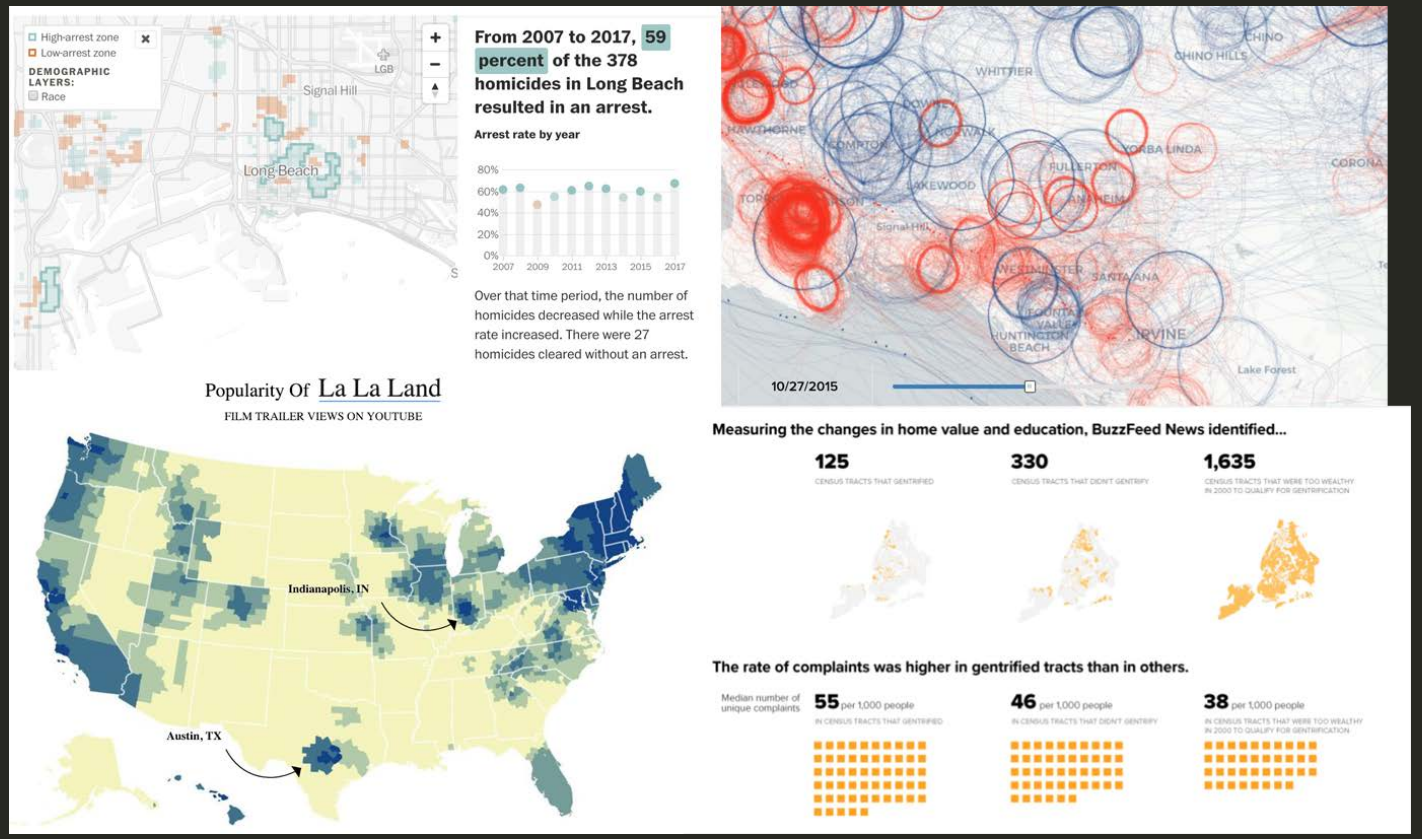

Actual vs predicted
Values predicted using a linear model all controls & department interaction
Male
Female
predicted
12.0
11.6
11.2
10.8
10.5
11.0
11.5
12.0
actual
gender
Male
Female

Advanced Data Journalism: Doing More with R

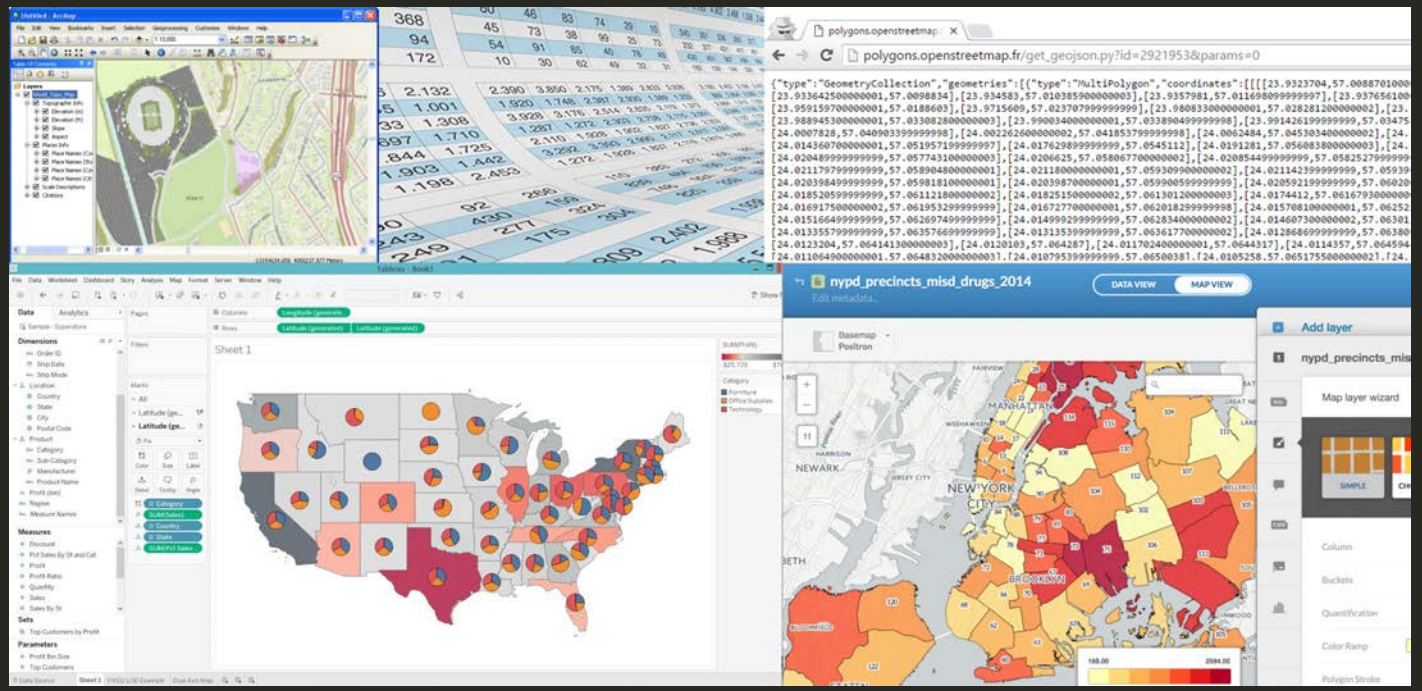
Module 4: Choropleth maps

Andrew Ba Tran

Maps are fun


High-arrest zone
Low-arrest zone
DEMOGRAPHIC LAYERS:
Race
LGB
Signal Hill
Long Beach
From 2007 to 2017, 59 percent of the 378 homicides in Long Beach resulted in an arrest.
Arrest rate by year
80%
60%
40%
20%
0%
2007 2009 2011 2013 2015 2017
Over that time period, the number of homicides decreased while the arrest rate increased. There were 27 homicides cleared without an arrest.
CHINO
CHINO HILLS
WHITTIER
HAWTHORNE
COMPTON
LAKEWOOD
FULLERTON
YORBA LINDA
CORONA
ANAHEIM
Signal Hill
SANTA ANA
HUNTINGTON BEACH
IRVINE
Lake Forest
10/27/2015
Popularity Of La La Land
FILM TRAILER VIEWS ON YOUTUBE
Indianapolis, IN
Austin, TX
Measuring the changes in home value and education, BuzzFeed News identified...
125
CENSUS TRACTS THAT GENTRIFIED
330
CENSUS TRACTS THAT DIDN'T GENTRIFY
1,635
CENSUS TRACTS THAT WERE TOO WEALTHY IN 2000 TO QUALIFY FOR GENTRIFICATION
The rate of complaints was higher in gentrified tracts than in others.
Median number of unique complaints
55 per 1,000 people
IN CENSUS TRACTS THAT GENTRIFIED
46 per 1,000 people
IN CENSUS TRACTS THAT DIDN'T GENTRIFY
38 per 1,000 people
IN CENSUS TRACTS THAT WERE TOO WEALTHY IN 2000 TO QUALIFY FOR GENTRIFICATION

Maps normally



Maps normally

1. **Download data and transform data**
    - Excel
2. **Find and download shapefiles**
    - Census TIGER
3. **Import maps and join with data and style**
    - ArcGIS or QGIS
4. **Export and tweak for style further**
    - Tableau, CartoDB, Illustrator

Mapping with R

Why map in R?

- Scripting and reproducibility
- Transparency and trust
- Easily interface with APIs for data and shapefiles
- Life is already complicated
    - Your process doesn't have to be

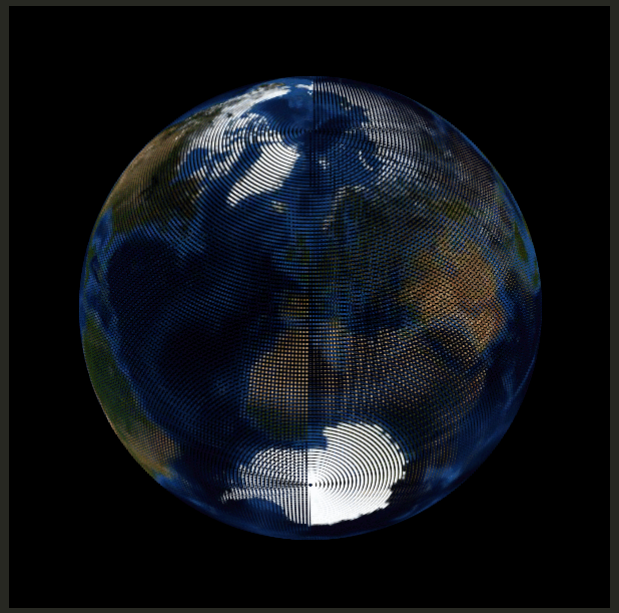
Basics

There are two underlying important pieces of information for spatial data:

- Coordinates of the object
- How the coordinates relate to a physical location on Earth
    - Also known as coordinate reference system or **CRS**

CRS

- Geographic
  - Uses three-dimensional model of the earth to define specific locations on the surface of the grid
  - longitude (East/West) and latitude (North/South)
- Projected
  - A translation of the three-dimensional grid onto a two-dimensional plane



Raster versus Vector data

Spatial data with a defined CRS can either be vector or raster data.

- Vector
  - Based on points that can be connected to form lines and polygons
  - Located with in a coordinate reference system
  - Example: Road map
- Raster
  - Are values within a grid system
  - Example: Satellite imagery

Shape files

Though we refer to a shape file in the singular, it's actually a collection of at least three basic files:

- .shp - lists shape and vertices
- .shx - has index with offsets
- .dbf - relationship file between geometry and attributes (data)

All files must be present in the directory and named the same (except for the file extension) to import correctly.

Let's load the packages we need:

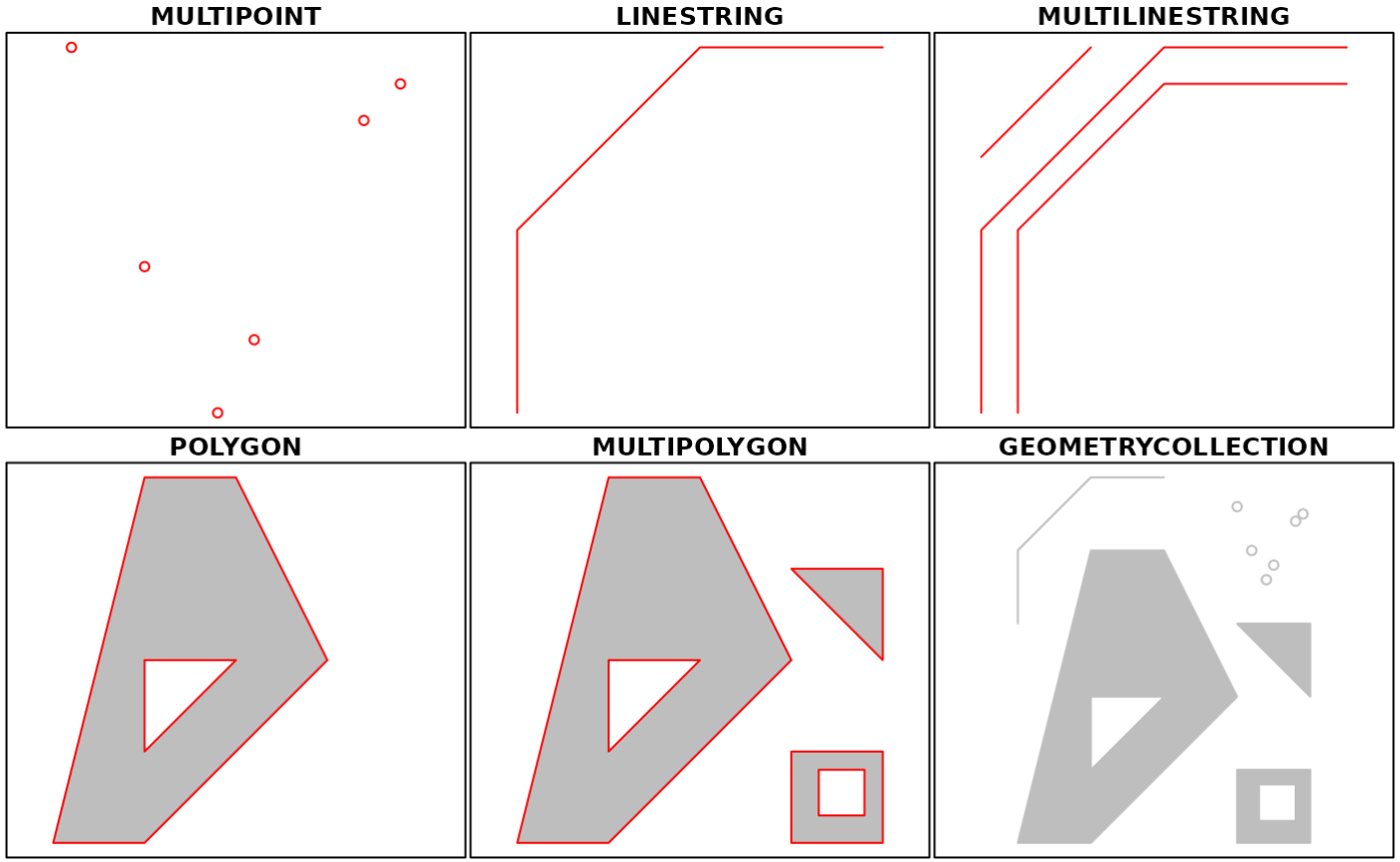
```
# Checking if the packages you need are installed -- if not, it will install for you
packages <- c("tidyverse", "stringr", "censusapi", "sf", "tidycensus", "ggspatial", "tigris")
if (length(setdiff(packages, rownames(installed.packages()))) > 0) {
  install.packages(setdiff(packages, rownames(installed.packages())), repos = "http://cran.us.r-project.org")
}

library(tidyverse)
library(sf)
library(ggspatial)
```

sf features

type	description
`POINT`	zero-dimensional geometry containing a single point
`LINESTRING`	sequence of points connected by straight, non-self intersecting line pieces; one-dimensional geometry
`POLYGON`	geometry with a positive area (two-dimensional); sequence of points form a closed, non-self intersecting ring; the first ring denotes the exterior ring, zero or more subsequent rings denote holes in this exterior ring
`MULTIPOINT`	set of points; a MULTIPOINT is simple if no two Points in the MULTIPOINT are equal
`MULTILINESTRING`	set of linestrings
`MULTIPOLYGON`	set of polygons
`GEOMETRYCOLLECTION`	set of geometries of any type except GEOMETRYCOLLECTION

sf features


MULTIPOINT
LINESTRING
MULTILINESTRING
POLYGON
MULTIPOLYGON
GEOMETRYCOLLECTION

Mapping a familiar shape file

st_read() is the function to import the shapefile.

Type out the code below or copy and paste it into the console or run Chunk1 from the XXXXX.rmd file

```
map_layer1 <- st_read("data/cases.shp")
```

```
Reading layer `cases' from data source
  `/Users/andrewtran/Documents/r_mooc_2022/data/cases.shp' using driver `ESRI Shapefile'
Simple feature collection with 250 features and 2 fields
Geometry type: POINT
Dimension:     XY
Bounding box:  xmin: -0.1400738 ymin: 51.51186 xmax: -0.1329335 ymax: 51.51583
Geodetic CRS:  WGS 84
```

```
glimpse(map_layer1)
```

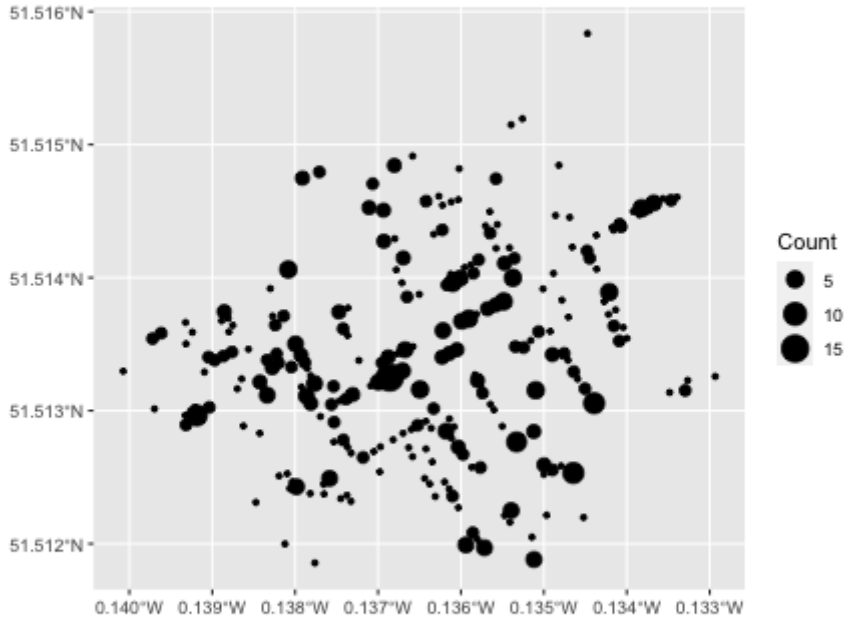
```
Rows: 250
Columns: 3
$ Id       <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
$ Count    <int> 3, 2, 1, 1, 4, 2, 2, 2, 3, 2, 2, 1, 3, 1, 4, 1, 1, 1, 4, 3, 2…
$ geometry <POINT [°]> POINT (-0.1379301 51.51342), POINT (-0.137883 51.51336)…
```

```
map_layer1
```

```
Simple feature collection with 250 features and 2 fields
Geometry type: POINT
Dimension:     XY
Bounding box:  xmin: -0.1400738 ymin: 51.51186 xmax: -0.1329335 ymax: 51.51583
Geodetic CRS:  WGS 84
First 10 features:
   Id Count                     geometry
1   0     3 POINT (-0.1379301 51.51342)
2   0     2  POINT (-0.137883 51.51336)
3   0     1 POINT (-0.1378529 51.51332)
4   0     1 POINT (-0.1378119 51.51326)
5   0     4  POINT (-0.1377668 51.5132)
6   0     2 POINT (-0.1375369 51.51318)
7   0     2 POINT (-0.1382004 51.51336)
8   0     2  POINT (-0.138045 51.51333)
9   0     3 POINT (-0.1382761 51.51332)
10  0     2 POINT (-0.1382234 51.51343)
```

```
map_layer1 %>%
  ggplot()
```

```
map_layer1 %>%
  ggplot() +
    geom_sf(aes(geometry=geometry, size=Count))
```



Does the pattern look familiar at all?

```
# bringing in another shape file
map_layer2 <- st_read("data/pumps.shp")
```

```
Reading layer `pumps' from data source
  `/Users/andrewtran/Documents/r_mooc_2022/data/pumps.shp' using driver `ESRI Shapefile'
Simple feature collection with 8 features and 1 field
Geometry type: POINT
Dimension:     XY
Bounding box:  xmin: -0.139671 ymin: 51.51002 xmax: -0.1316298 ymax: 51.51491
Geodetic CRS:  WGS 84
```

Does the pattern look familiar at all?

```
# bringing in another shape file
map_layer2 <- st_read("data/pumps.shp")

map_layer1
```

```
Reading layer `pumps' from data source
  `/Users/andrewtran/Documents/r_mooc_2022/data/pumps.shp' using driver `ESRI Shapefile'
Simple feature collection with 8 features and 1 field
Geometry type: POINT
Dimension:     XY
Bounding box:  xmin: -0.139671 ymin: 51.51002 xmax: -0.1316298 ymax: 51.51491
Geodetic CRS:  WGS 84

Simple feature collection with 250 features and 2 fields
Geometry type: POINT
Dimension:     XY
Bounding box:  xmin: -0.1400738 ymin: 51.51186 xmax: -0.1329335 ymax: 51.51583
Geodetic CRS:  WGS 84
First 10 features:
   Id Count                    geometry
1   0     3 POINT (-0.1379301 51.51342)
2   0     2  POINT (-0.137883 51.51336)
3   0     1 POINT (-0.1378529 51.51332)
4   0     1 POINT (-0.1378119 51.51326)
5   0     4  POINT (-0.1377668 51.5132)
6   0     2 POINT (-0.1375369 51.51318)
7   0     2 POINT (-0.1382004 51.51336)
8   0     2  POINT (-0.138045 51.51333)
9   0     3 POINT (-0.1382761 51.51332)
10  0     2 POINT (-0.1382234 51.51343)
```

Does the pattern look familiar at all?

```
# bringing in another shape file
map_layer2 <- st_read("data/pumps.shp")

map_layer1 %>%
ggplot()
```

```
Reading layer `pumps' from data source
  `/Users/andrewtran/Documents/r_mooc_2022/data/pumps.shp' using driver `ESRI Shapefile'
Simple feature collection with 8 features and 1 field
Geometry type: POINT
Dimension:     XY
Bounding box:  xmin: -0.139671 ymin: 51.51002 xmax: -0.1316298 ymax: 51.51491
Geodetic CRS:  WGS 84
```

Does the pattern look familiar at all?

```
# bringing in another shape file
map_layer2 <- st_read("data/pumps.shp")

map_layer1 %>%
ggplot() +
  geom_sf(aes(geometry=geometry, size=Count))
```

```
Reading layer `pumps' from data source
  `/Users/andrewtran/Documents/r_mooc_2022/data/pumps.shp' using driver `ESRI Shapefile'
Simple feature collection with 8 features and 1 field
Geometry type: POINT
Dimension:     XY
Bounding box:  xmin: -0.139671 ymin: 51.51002 xmax: -0.1316298 ymax: 51.51491
Geodetic CRS:  WGS 84
```

Does the pattern look familiar at all?

```
# bringing in another shape file
map_layer2 <- st_read("data/pumps.shp")

map_layer1 %>%
ggplot() +
  geom_sf(aes(geometry=geometry, size=Count)) +
  geom_sf(data=map_layer2, aes(size = 3, color = "re
```

```
Reading layer `pumps' from data source
  `/Users/andrewtran/Documents/r_mooc_2022/data/pumps.shp' using driver `ESRI Shapefile'
Simple feature collection with 8 features and 1 field
Geometry type: POINT
Dimension:     XY
Bounding box:  xmin: -0.139671 ymin: 51.51002 xmax: -0.1316298 ymax: 51.51491
Geodetic CRS:  WGS 84
```

```
map_layer1 %>%
  ggplot() +
    annotation_map_tile(type = "osm", zoomin = 0) +
    geom_sf(aes(geometry=geometry, size=Count), alpha = 0.7) +
    geom_sf(data=map_layer2, aes(size = 3, color = "red")) +
    theme_void()
```

This is the hypothetical data John Snow was working with.

Date	Last_Name	First_Name	Address	Age	Cause_death
Aug 31, 1854	Jones	Thomas	26 Broad St.	37	cholera
Aug 31, 1854	Jones	Mary	26 Broad St.	11	cholera
Oct 1, 1854	Warwick	Martin	14 Broad St.	23	cholera

Mapping data

With sf and ggplot and tigris

tigris package functions

Function	Datasets available	Years available
`nation()`	cartographic (1:5m; 1:20m)	2013-2021
`divisions()`	cartographic (1:500k; 1:5m; 1:20m)	2013-2021
`regions()`	cartographic (1:500k; 1:5m; 1:20m)	2013-2021
`states()`	TIGER/Line; cartographic (1:500k; 1:5m; 1:20m)	1990, 2000, 2010-2021
`counties()`	TIGER/Line; cartographic (1:500k; 1:5m; 1:20m)	1990, 2000, 2010-2021
`tracts()`	TIGER/Line; cartographic (1:500k)	1990, 2000, 2010-2021
`block_groups()`	TIGER/Line; cartographic (1:500k)	1990, 2000, 2010-2021
`blocks()`	TIGER/Line	2000, 2010-2021
`places()`	TIGER/Line; cartographic (1:500k)	2011-2021
`pumas()`	TIGER/Line; cartographic (1:500k)	2012-2021
`school_districts()`	TIGER/Line; cartographic	2011-2021

Read the documentation

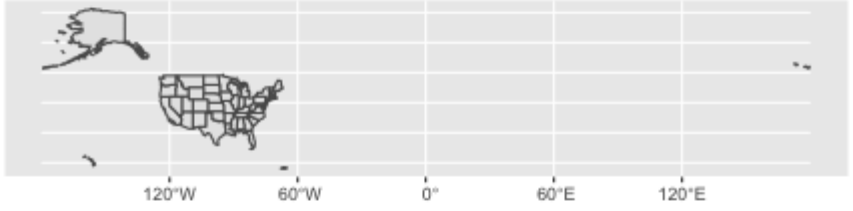
```r
library(tigris)
us_states <- states(cb = TRUE, resolution = "20m")
```

```r
library(tigris)
us_states <- states(cb = TRUE, resolution = "20m")
```

```r
glimpse(us_states)
```

```
Rows: 52
Columns: 10
$ STATEFP  <chr> "47", "27", "17", "30", "11", "06", "21", "10", "48", "55", "…
$ STATENS  <chr> "01325873", "00662849", "01779784", "00767982", "01702382", "…
$ AFFGEOID <chr> "0400000US47", "0400000US27", "0400000US17", "0400000US30", "…
$ GEOID    <chr> "47", "27", "17", "30", "11", "06", "21", "10", "48", "55", "…
$ STUSPS   <chr> "TN", "MN", "IL", "MT", "DC", "CA", "KY", "DE", "TX", "WI", "…
$ NAME     <chr> "Tennessee", "Minnesota", "Illinois", "Montana", "District of…
$ LSAD     <chr> "00", "00", "00", "00", "00", "00", "00", "00", "00", "00", "…
$ ALAND    <dbl> 1.067916e+11, 2.062322e+11, 1.437785e+11, 3.769737e+11, 1.583…
$ AWATER   <dbl> 2322913374, 18949864226, 6216594318, 3866689601, 18709762, 20…
$ geometry <MULTIPOLYGON [°]> MULTIPOLYGON (((-90.3007 35..., MULTIPOLYGON (((…
```

```
us_states %>%
  ggplot() +
  geom_sf()
```



```r
us_states <- states(cb = TRUE, resolution = "20m") %>%
     shift_geometry()

us_states %>%
  ggplot() +
  geom_sf()
```

Styling maps

```
us_states %>%
  ggplot() +
  geom_sf(color="red") +
  theme_void()
```

Join data to the shapefiles

```
library(jsonlite)
```

```
library(jsonlite)

fl_opioids <- fromJSON("https://arcos-api.ext.nile.w
```

```r
library(jsonlite)

fl_opioids <- fromJSON("https://arcos-api.ext.nile.w

glimpse(fl_opioids)
```

```
Rows: 602
Columns: 6
$ BUYER_COUNTY <chr> "ALACHUA", "ALACHUA", "ALACHUA", "ALACHUA", "ALACHUA", "A…
$ BUYER_STATE  <chr> "FL", "FL", "FL", "FL", "FL", "FL", "FL", "FL", "FL", "FL…
$ year         <int> 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 200…
$ count        <int> 20923, 21998, 23579, 24119, 23624, 23446, 22085, 22662, 2…
$ DOSAGE_UNIT  <dbl> 7029756, 7849764, 8786119, 9820973, 9760670, 9622669, 862…
$ countyfips   <chr> "12001", "12001", "12001", "12001", "12001", "12001", "12…
```

```r
library(jsonlite)

fl_opioids <- fromJSON("https://arcos-api.ext.nile.w

glimpse(fl_opioids)

fl_pop <- fromJSON("https://arcos-api.ext.nile.works
```

```
Rows: 602
Columns: 6
$ BUYER_COUNTY <chr> "ALACHUA", "ALACHUA", "ALACHUA", "ALACHUA", "ALACHUA", "A…
$ BUYER_STATE  <chr> "FL", "FL", "FL", "FL", "FL", "FL", "FL", "FL", "FL", "FL…
$ year         <int> 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 200…
$ count        <int> 20923, 21998, 23579, 24119, 23624, 23446, 22085, 22662, 2…
$ DOSAGE_UNIT  <dbl> 7029756, 7849764, 8786119, 9820973, 9760670, 9622669, 862…
$ countyfips   <chr> "12001", "12001", "12001", "12001", "12001", "12001", "12…
```

```r
library(jsonlite)

fl_opioids <- fromJSON("https://arcos-api.ext.nile.w

glimpse(fl_opioids)

fl_pop <- fromJSON("https://arcos-api.ext.nile.works

glimpse(fl_pop)
```

```
Rows: 602
Columns: 6
$ BUYER_COUNTY <chr> "ALACHUA", "ALACHUA", "ALACHUA", "ALACHUA", "ALACHUA", "A…
$ BUYER_STATE  <chr> "FL", "FL", "FL", "FL", "FL", "FL", "FL", "FL", "FL", "FL…
$ year         <int> 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 200…
$ count        <int> 20923, 21998, 23579, 24119, 23624, 23446, 22085, 22662, 2…
$ DOSAGE_UNIT  <dbl> 7029756, 7849764, 8786119, 9820973, 9760670, 9622669, 862…
$ countyfips   <chr> "12001", "12001", "12001", "12001", "12001", "12001", "12…

Rows: 603
Columns: 10
$ BUYER_COUNTY <chr> "ALACHUA", "BAKER", "BAY", "BRADFORD", "BREVARD", "BROWAR…
$ BUYER_STATE  <chr> "FL", "FL", "FL", "FL", "FL", "FL", "FL", "FL", "FL", "FL…
$ countyfips   <chr> "12001", "12003", "12005", "12007", "12009", "12011", "12…
$ STATE        <int> 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 1…
$ COUNTY       <int> 1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 27, 29, 31, 33…
$ county_name  <chr> "Alachua", "Baker", "Bay", "Bradford", "Brevard", "Browar…
$ NAME         <chr> "Alachua County, Florida", "Baker County, Florida", "Bay …
$ variable     <chr> "B01003_001", "B01003_001", "B01003_001", "B01003_001", "…
$ year         <int> 2006, 2006, 2006, 2006, 2006, 2006, 2006, 2006, 2006, 200…
$ population   <int> 239506, 25571, 165644, 28506, 535138, 1739348, 13997, 157…
```

```
fl_data <- left_join(fl_pop, fl_opioids)
```

```r
fl_data <- left_join(fl_pop, fl_opioids)

glimpse(fl_data)
```

```
Rows: 603
Columns: 12
$ BUYER_COUNTY <chr> "ALACHUA", "BAKER", "BAY", "BRADFORD", "BREVARD", "BROWAR…
$ BUYER_STATE  <chr> "FL", "FL", "FL", "FL", "FL", "FL", "FL", "FL", "FL", "FL…
$ countyfips   <chr> "12001", "12003", "12005", "12007", "12009", "12011", "12…
$ STATE        <int> 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 1…
$ COUNTY       <int> 1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 27, 29, 31, 33…
$ county_name  <chr> "Alachua", "Baker", "Bay", "Bradford", "Brevard", "Browar…
$ NAME         <chr> "Alachua County, Florida", "Baker County, Florida", "Bay …
$ variable     <chr> "B01003_001", "B01003_001", "B01003_001", "B01003_001", "…
$ year         <int> 2006, 2006, 2006, 2006, 2006, 2006, 2006, 2006, 2006, 200…
$ population   <int> 239506, 25571, 165644, 28506, 535138, 1739348, 13997, 157…
$ count        <int> 20923, 2922, 20910, 3489, 57900, 132503, 1061, 17498, 156…
$ DOSAGE_UNIT  <dbl> 7029756, 1382600, 8415570, 1280200, 19751285, 58919069, 2…
```

```r
fl_data <- left_join(fl_pop, fl_opioids)

glimpse(fl_data)

fl_data <- fl_data
```

```
Rows: 603
Columns: 12
$ BUYER_COUNTY <chr> "ALACHUA", "BAKER", "BAY", "BRADFORD", "BREVARD", "BROWAR…
$ BUYER_STATE  <chr> "FL", "FL", "FL", "FL", "FL", "FL", "FL", "FL", "FL", "FL…
$ countyfips   <chr> "12001", "12003", "12005", "12007", "12009", "12011", "12…
$ STATE        <int> 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 1…
$ COUNTY       <int> 1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 27, 29, 31, 33…
$ county_name  <chr> "Alachua", "Baker", "Bay", "Bradford", "Brevard", "Browar…
$ NAME         <chr> "Alachua County, Florida", "Baker County, Florida", "Bay …
$ variable     <chr> "B01003_001", "B01003_001", "B01003_001", "B01003_001", "…
$ year         <int> 2006, 2006, 2006, 2006, 2006, 2006, 2006, 2006, 2006, 200…
$ population   <int> 239506, 25571, 165644, 28506, 535138, 1739348, 13997, 157…
$ count        <int> 20923, 2922, 20910, 3489, 57900, 132503, 1061, 17498, 156…
$ DOSAGE_UNIT  <dbl> 7029756, 1382600, 8415570, 1280200, 19751285, 58919069, 2…
```

```
fl_data <- left_join(fl_pop, fl_opioids)

glimpse(fl_data)

fl_data <- fl_data %>%
  mutate(dosage_per_person=round(DOSAGE_UNIT/populat
```

```
Rows: 603
Columns: 12
$ BUYER_COUNTY <chr> "ALACHUA", "BAKER", "BAY", "BRADFORD", "BREVARD", "BROWAR…
$ BUYER_STATE  <chr> "FL", "FL", "FL", "FL", "FL", "FL", "FL", "FL", "FL", "FL…
$ countyfips   <chr> "12001", "12003", "12005", "12007", "12009", "12011", "12…
$ STATE        <int> 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 1…
$ COUNTY       <int> 1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 27, 29, 31, 33…
$ county_name  <chr> "Alachua", "Baker", "Bay", "Bradford", "Brevard", "Browar…
$ NAME         <chr> "Alachua County, Florida", "Baker County, Florida", "Bay …
$ variable     <chr> "B01003_001", "B01003_001", "B01003_001", "B01003_001", "…
$ year         <int> 2006, 2006, 2006, 2006, 2006, 2006, 2006, 2006, 2006, 200…
$ population   <int> 239506, 25571, 165644, 28506, 535138, 1739348, 13997, 157…
$ count        <int> 20923, 2922, 20910, 3489, 57900, 132503, 1061, 17498, 156…
$ DOSAGE_UNIT  <dbl> 7029756, 1382600, 8415570, 1280200, 19751285, 58919069, 2…
```

```r
fl_data <- left_join(fl_pop, fl_opioids)

glimpse(fl_data)

fl_data <- fl_data %>%
  mutate(dosage_per_person=round(DOSAGE_UNIT/populat

glimpse(fl_data)
```

```
Rows: 603
Columns: 12
$ BUYER_COUNTY <chr> "ALACHUA", "BAKER", "BAY", "BRADFORD", "BREVARD", "BROWAR…
$ BUYER_STATE  <chr> "FL", "FL", "FL", "FL", "FL", "FL", "FL", "FL", "FL", "FL…
$ countyfips   <chr> "12001", "12003", "12005", "12007", "12009", "12011", "12…
$ STATE        <int> 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 1…
$ COUNTY       <int> 1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 27, 29, 31, 33…
$ county_name  <chr> "Alachua", "Baker", "Bay", "Bradford", "Brevard", "Browar…
$ NAME         <chr> "Alachua County, Florida", "Baker County, Florida", "Bay …
$ variable     <chr> "B01003_001", "B01003_001", "B01003_001", "B01003_001", "…
$ year         <int> 2006, 2006, 2006, 2006, 2006, 2006, 2006, 2006, 2006, 200…
$ population   <int> 239506, 25571, 165644, 28506, 535138, 1739348, 13997, 157…
$ count        <int> 20923, 2922, 20910, 3489, 57900, 132503, 1061, 17498, 156…
$ DOSAGE_UNIT  <dbl> 7029756, 1382600, 8415570, 1280200, 19751285, 58919069, 2…

Rows: 603
Columns: 13
$ BUYER_COUNTY      <chr> "ALACHUA", "BAKER", "BAY", "BRADFORD", "BREVARD", "B…
$ BUYER_STATE       <chr> "FL", "FL", "FL", "FL", "FL", "FL", "FL", "FL", "FL"…
$ countyfips        <chr> "12001", "12003", "12005", "12007", "12009", "12011"…
$ STATE             <int> 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, …
$ COUNTY            <int> 1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 27, 29, 3…
$ county_name       <chr> "Alachua", "Baker", "Bay", "Bradford", "Brevard", "B…
$ NAME              <chr> "Alachua County, Florida", "Baker County, Florida", …
$ variable          <chr> "B01003_001", "B01003_001", "B01003_001", "B01003_00…
$ year              <int> 2006, 2006, 2006, 2006, 2006, 2006, 2006, 2006, 2006…
$ population        <int> 239506, 25571, 165644, 28506, 535138, 1739348, 13997…
$ count             <int> 20923, 2922, 20910, 3489, 57900, 132503, 1061, 17498…
$ DOSAGE_UNIT       <dbl> 7029756, 1382600, 8415570, 1280200, 19751285, 589190…
$ dosage_per_person <dbl> 29.4, 54.1, 50.8, 44.9, 36.9, 33.9, 20.6, 39.8, 41.3…
```

```
## Function to download county shapefiles in Florida
fl_shape <- counties(state="FL", cb=T)
```

```
|
|                                                                      |   0%
|
|                                                                      |   1%
|
|=                                                                     |   1%
|
|=                                                                     |   2%
|
|==                                                                    |   2%
|
|==                                                                    |   3%
|
|===                                                                   |   4%
|
|===                                                                   |   5%
|
|====                                                                  |   5%
|
|====                                                                  |   6%
|
|=====                                                                 |   7%
|
|=====                                                                 |   8%
|
|======                                                                |   8%
|
|======                                                                |   9%
|
|=======                                                               |  10%
|
|=======                                                               |  11%
```

```r
## Function to download county shapefiles in Florida
fl_shape <- counties(state="FL", cb=T)

glimpse(fl_shape)
```

```
Rows: 67
Columns: 13
$ STATEFP    <chr> "12", "12", "12", "12", "12", "12", "12", "12", "12", "12",…
$ COUNTYFP   <chr> "009", "101", "037", "053", "045", "047", "095", "083", "10…
$ COUNTYNS   <chr> "00295749", "00295739", "00306911", "00295751", "00306917",…
$ AFFGEOID   <chr> "0500000US12009", "0500000US12101", "0500000US12037", "0500…
$ GEOID      <chr> "12009", "12101", "12037", "12053", "12045", "12047", "1209…
$ NAME       <chr> "Brevard", "Pasco", "Franklin", "Hernando", "Gulf", "Hamilt…
$ NAMELSAD   <chr> "Brevard County", "Pasco County", "Franklin County", "Herna…
$ STUSPS     <chr> "FL", "FL", "FL", "FL", "FL", "FL", "FL", "FL", "FL", "FL",…
$ STATE_NAME <chr> "Florida", "Florida", "Florida", "Florida", "Florida", "Flo…
$ LSAD       <chr> "06", "06", "06", "06", "06", "06", "06", "06", "06", "06",…
$ ALAND      <dbl> 2628762626, 1933733392, 1411498965, 1224975810, 1433437353,…
$ AWATER     <dbl> 1403940953, 694477432, 2270440522, 627928028, 624436316, 12…
$ geometry   <MULTIPOLYGON [°]> MULTIPOLYGON (((-80.98725 2..., MULTIPOLYGON (…
```

```r
## Function to download county shapefiles in Florida
fl_shape <- counties(state="FL", cb=T)

glimpse(fl_shape)

fl <- left_join(fl_shape, fl_data, by=c("GEOID"="cou
```

```
Rows: 67
Columns: 13
$ STATEFP    <chr> "12", "12", "12", "12", "12", "12", "12", "12", "12", "12",…
$ COUNTYFP   <chr> "009", "101", "037", "053", "045", "047", "095", "083", "10…
$ COUNTYNS   <chr> "00295749", "00295739", "00306911", "00295751", "00306917",…
$ AFFGEOID   <chr> "0500000US12009", "0500000US12101", "0500000US12037", "0500…
$ GEOID      <chr> "12009", "12101", "12037", "12053", "12045", "12047", "1209…
$ NAME       <chr> "Brevard", "Pasco", "Franklin", "Hernando", "Gulf", "Hamilt…
$ NAMELSAD   <chr> "Brevard County", "Pasco County", "Franklin County", "Herna…
$ STUSPS     <chr> "FL", "FL", "FL", "FL", "FL", "FL", "FL", "FL", "FL", "FL",…
$ STATE_NAME <chr> "Florida", "Florida", "Florida", "Florida", "Florida", "Flo…
$ LSAD       <chr> "06", "06", "06", "06", "06", "06", "06", "06", "06", "06",…
$ ALAND      <dbl> 2628762626, 1933733392, 1411498965, 1224975810, 1433437353,…
$ AWATER     <dbl> 1403940953, 694477432, 2270440522, 627928028, 624436316, 12…
$ geometry   <MULTIPOLYGON [°]> MULTIPOLYGON (((-80.98725 2..., MULTIPOLYGON (…
```

```
glimpse(fl)
```

```
Rows: 603
Columns: 25
$ STATEFP           <chr> "12", "12", "12", "12", "12", "12", "12", "12", "12"…
$ COUNTYFP          <chr> "009", "009", "009", "009", "009", "009", "009", "00…
$ COUNTYNS          <chr> "00295749", "00295749", "00295749", "00295749", "002…
$ AFFGEOID          <chr> "0500000US12009", "0500000US12009", "0500000US12009"…
$ GEOID             <chr> "12009", "12009", "12009", "12009", "12009", "12009"…
$ NAME.x            <chr> "Brevard", "Brevard", "Brevard", "Brevard", "Brevard…
$ NAMELSAD          <chr> "Brevard County", "Brevard County", "Brevard County"…
$ STUSPS            <chr> "FL", "FL", "FL", "FL", "FL", "FL", "FL", "FL", "FL"…
$ STATE_NAME        <chr> "Florida", "Florida", "Florida", "Florida", "Florida…
$ LSAD              <chr> "06", "06", "06", "06", "06", "06", "06", "06", "06"…
$ ALAND             <dbl> 2628762626, 2628762626, 2628762626, 2628762626, 2628…
$ AWATER            <dbl> 1403940953, 1403940953, 1403940953, 1403940953, 1403…
$ BUYER_COUNTY      <chr> "BREVARD", "BREVARD", "BREVARD", "BREVARD", "BREVARD…
$ BUYER_STATE       <chr> "FL", "FL", "FL", "FL", "FL", "FL", "FL", "FL", "FL"…
$ STATE             <int> 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, …
$ COUNTY            <int> 9, 9, 9, 9, 9, 9, 9, 9, 9, 101, 101, 101, 101, 101, …
$ county_name       <chr> "Brevard", "Brevard", "Brevard", "Brevard", "Brevard…
$ NAME.y            <chr> "Brevard County, Florida", "Brevard County, Florida"…
$ variable          <chr> "B01003_001", "B01003_001", "B01003_001", "B01003_00…
$ year              <int> 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014…
$ population        <int> 535138, 539719, 542378, 532697, 540583, 542320, 5440…
$ count             <int> 57900, 62911, 64517, 61685, 64609, 66856, 66605, 687…
$ DOSAGE_UNIT       <dbl> 19751285, 22323482, 25587640, 27191952, 31189425, 31…
$ dosage_per_person <dbl> 36.9, 41.4, 47.2, 51.0, 57.7, 57.2, 49.4, 46.8, 47.6…
$ geometry          <MULTIPOLYGON [°]> MULTIPOLYGON (((-80.98725 2..., MULTIPO…
```

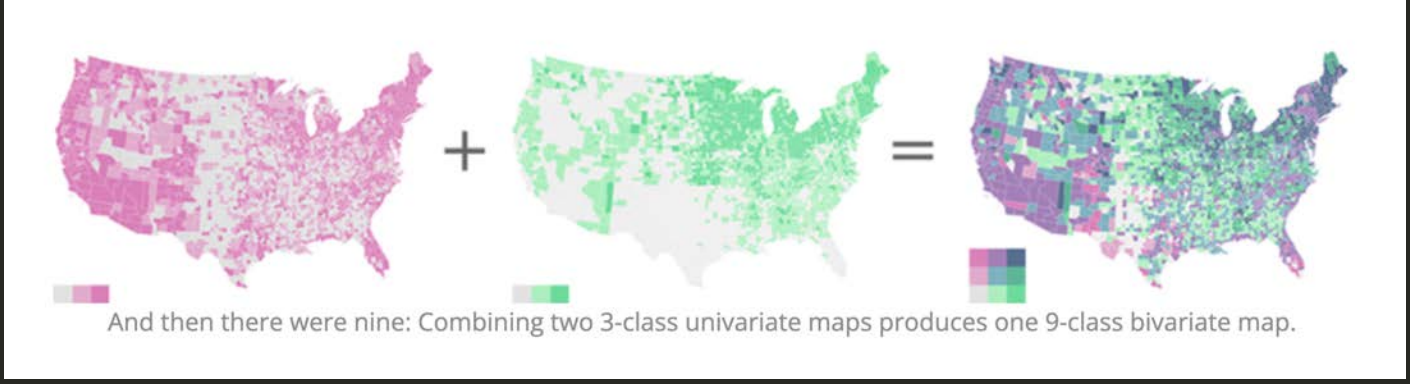
```
fl %>%
  filter(year==2014) %>%
  ggplot() +
  geom_sf(aes(geometry=geometry,
              fill = dosage_per_person,
              color = dosage_per_person))
```

Small multiples and styling

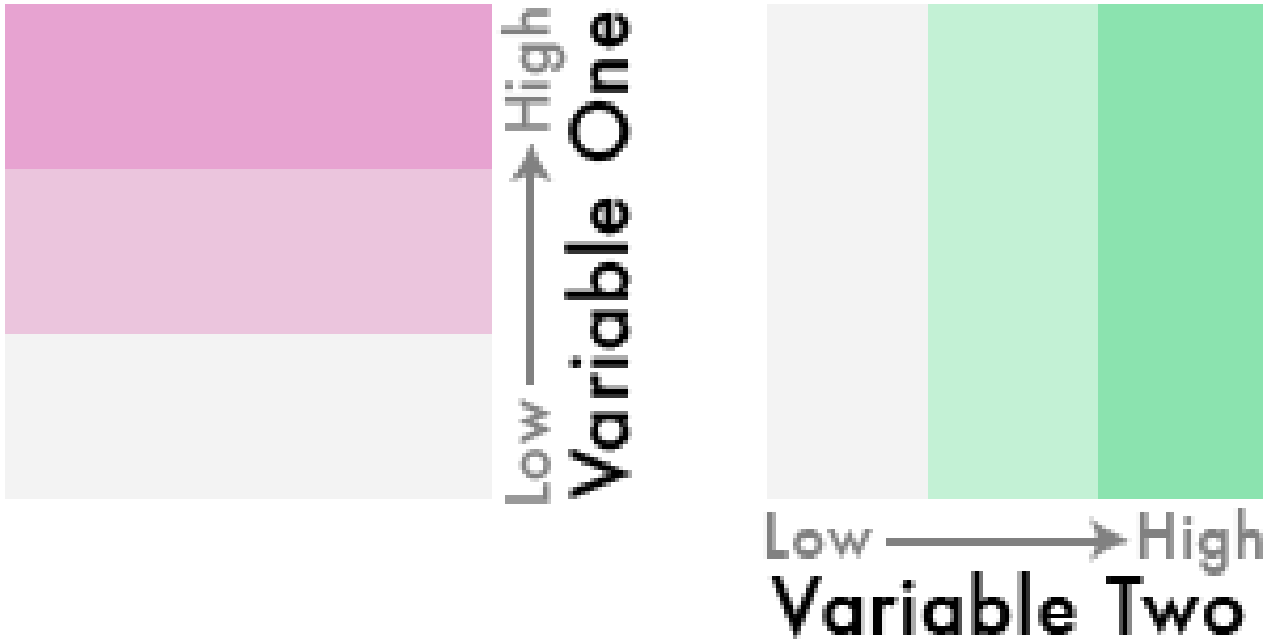
```r
library(viridis)
fl %>%
  ggplot() +
  geom_sf(aes(geometry=geometry, fill = dosage_per_person), color=NA) +
  facet_wrap(~year, ncol=4) +
  scale_fill_viridis(direction=-1) +
  theme_void() +
  labs(title="Oxycodone and hydrocodone pills in Florida", caption="Source: The Washington Post, ARCOS")
```

Bivariate maps

In the exercises



And then there were nine: Combining two 3-class univariate maps produces one 9-class bivariate map.


High
Low
Variable One
Low
High
Variable Two

Save it with cowplot

```
library(cowplot)

save_plot("name_of_file.png", ggplot_object, base_height = NULL, base_width = 12)

#for svgs,
#install.packages("svglite") to make this work
save_plot("name_of_file.svg", ggplot_object, base_height = NULL, base_width = 12)

#as a shapefile?
st_write(ggplot_object, "name_of_file.geojson")
st_write(ggplot_object, "name_of_file.shp")
```