PLSC476: Empirical Legal Studies

Christopher Zorn

January 28, 2021

R and RMarkdown

- RMarkdown is "an authoring framework for data science" that facilitates literate programming and reproduceable research
- Purpose: To allow for the creation of *reproduceable*, *dynamic* documents, presentations, and web pages.
- Similar to Markdown (and other lightweight, literate markup languages)
- Combines text, `R` code, and `R` output
- Outut formats: PDF, HTML, Word, Shiny (for interactive web apps), others
- These slides were created using RMarkdown; the code for doing so is available on the course github repository

Why Plot?

- Know your data

- Catch mistakes

- Learn something. . .

Running Example: COVID-19 in PA

- $N = 67$ counties
- Current as of yesterday (1/27/2021)
- Source: PA Department of Health

Variables:

- County Name
- Latitude/Longitude
- FIPS code
- Date
- New Cases
- 7-Day Average of New Cases
- Total/Cumulative Cases
- County Population

- New Case Rate (per 100K population)
- 7-Day Average of New Case Rate
- Cumulative Case Rate (per 100K population)
- New Deaths
- Total/Cumulative Deaths
- Cumulative Death Rate (per 100K population)

Table 1: PA COVID-19 Data, 1/27/2021

	FIPS	Date	County	New.Cases	New.Cases.7dayAvg	Cumulative.Cases
360	42001	01/27/2021	Adams	47	50.4	6446
693	42003	01/27/2021	Allegheny	350	344.0	68445
1026	42005	01/27/2021	Armstrong	21	20.7	4560
1359	42007	01/27/2021	Beaver	68	52.0	10833
1692	42009	01/27/2021	Bedford	13	18.7	3587
2025	42011	01/27/2021	Berks	185	198.0	30822
2358	42013	01/27/2021	Blair	31	41.0	9817
2691	42015	01/27/2021	Bradford	33	25.6	3866
3024	42017	01/27/2021	Bucks	279	242.6	38977
3357	42019	01/27/2021	Butler	77	74.0	12566
3690	42021	01/27/2021	Cambria	30	44.6	10664
4023	42023	01/27/2021	Cameron	0	2.3	237

Data Summary

```
      FIPS            Date              County            New.Cases
 Min.   :42001   Length:67          Length:67          Min.   :  0
 1st Qu.:42034   Class :character   Class :character   1st Qu.: 13
 Median :42067   Mode  :character   Mode  :character   Median : 33
 Mean   :42067                                         Mean   : 86
 3rd Qu.:42100                                         3rd Qu.: 75
 Max.   :42133                                         Max.   :571
 New.Cases.7dayAvg Cumulative.Cases  Population        New.Case.Rate
 Min.   :  1       Min.   :   237   Min.   :   4447   Min.   :  0.0
 1st Qu.: 14       1st Qu.:  2472   1st Qu.:  42025   1st Qu.: 26.8
 Median : 37       Median :  5846   Median :  84629   Median : 35.2
 Mean   : 75       Mean   : 12305   Mean   : 191074   Mean   : 44.6
 3rd Qu.: 86       3rd Qu.: 12457   3rd Qu.: 208270   3rd Qu.: 44.1
 Max.   :420       Max.   :102870   Max.   :1584064   Max.   :194.5
 New.Case.Rate.7DayAvg Cumulative.Case.Rate  Longitude        Latitude
 Min.   : 11.8         Min.   : 3675        Min.   :-80.4   Min.   :39.9
 1st Qu.: 31.7         1st Qu.: 5689        1st Qu.:-79.1   1st Qu.:40.4
 Median : 37.9         Median : 6594        Median :-77.4   Median :40.8
 Mean   : 41.0         Mean   : 6655        Mean   :-77.6   Mean   :40.8
 3rd Qu.: 49.7         3rd Qu.: 7190        3rd Qu.:-76.2   3rd Qu.:41.3
 Max.   :112.2         Max.   :17262        Max.   :-75.0   Max.   :42.0
 Georeferenced.Lat...Long  New.Deaths      Total.Deaths    Total.Death.Rate
 Length:67                Min.   :0.000   Min.   :   7   Min.   : 79
 Class :character         1st Qu.:0.000   1st Qu.:  72   1st Qu.:138
 Mode  :character         Median :0.000   Median : 151   Median :168
                          Mean   :0.015   Mean   : 318   Mean   :173
                          3rd Qu.:0.000   3rd Qu.: 345   3rd Qu.:191
                          Max.   :1.000   Max.   :2879   Max.   :327
```

Univariate Graphics

- Dotcharts / Barcharts
- Histograms
- Density Plots
- Boxplots
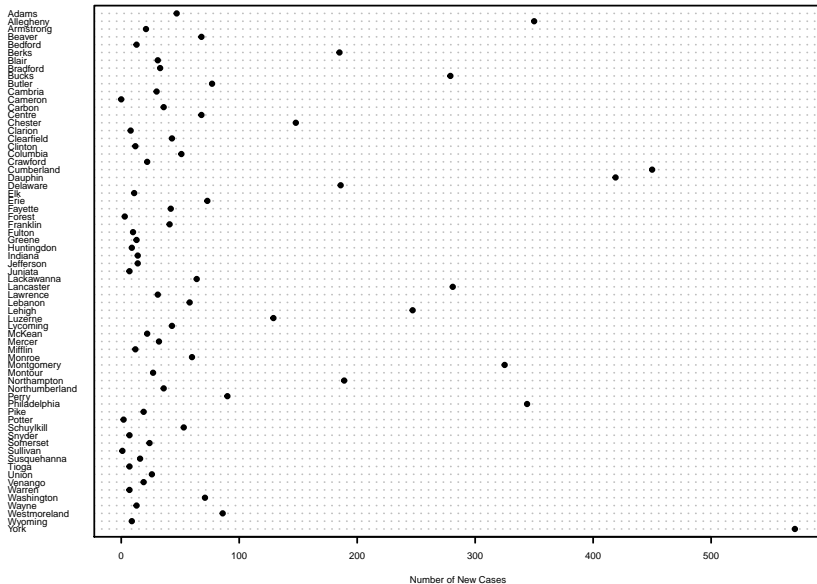- Q-Q Plots
- Others. . .

A Dotchart

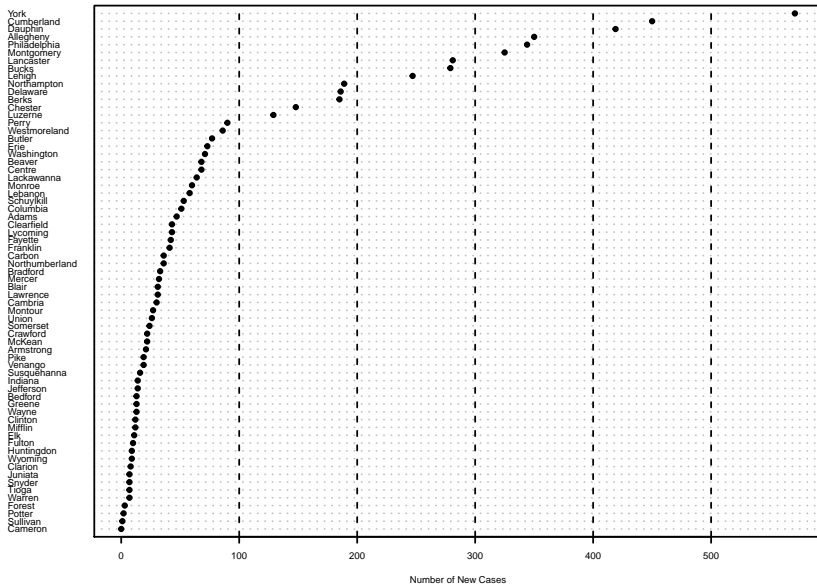Figure 1: New COVID-19 Cases by County

A Sorted Dotchart

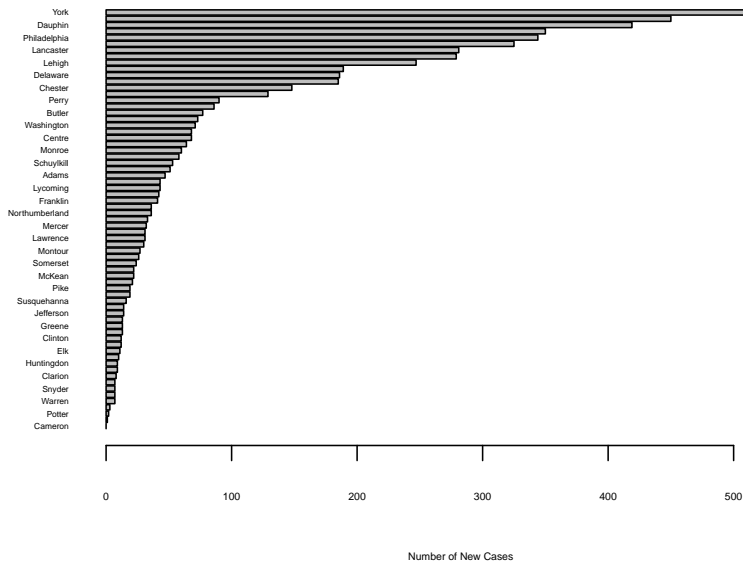Figure 2: New COVID-19 Cases by County

A (Sorted) Barchart

Figure 3: New COVID-19 Cases by County

The Histogram: Cumulative Case Percentages

```
   vars  n mean   sd median trimmed  mad  min  max range skew kurtosis   se
X1    1 67 6.66 1.76   6.59    6.54 1.11 3.67 17.3  13.6 3.09      17.3 0.22
```


COVID−19 Infection Percent
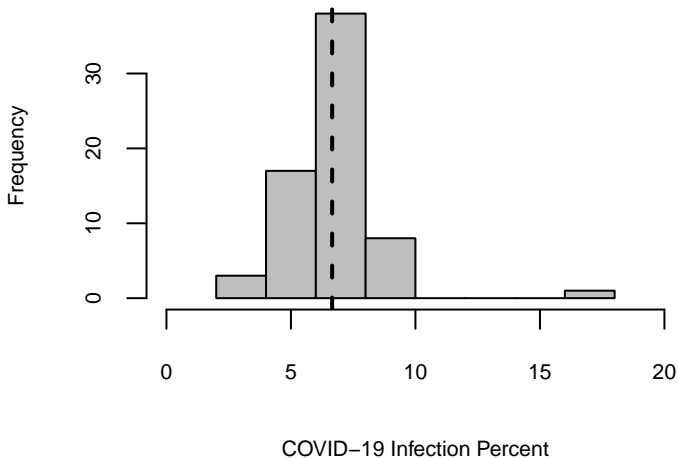
Figure 4: Cumulative Cases, as a Percentage of Population

"Kernel Density" Plot

COVID−19 Infection Percent
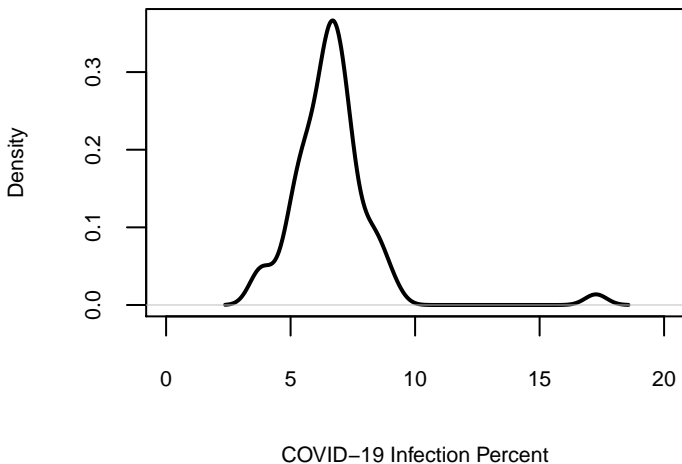
Figure 5: Cumulative Cases, as a Percentage of Population

Density + Histogram

COVID−19 Infection Percent
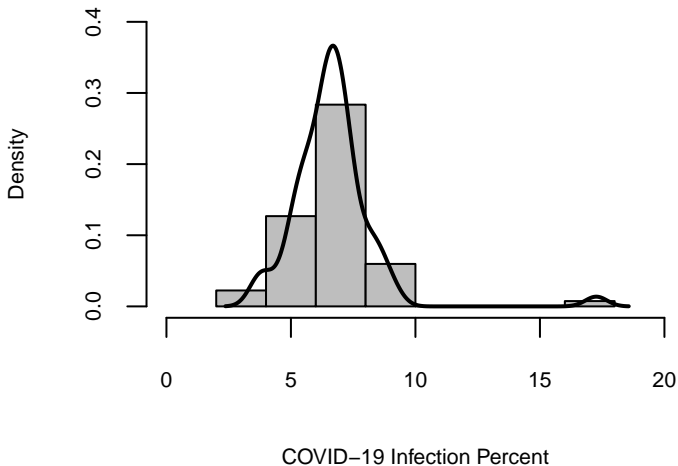
Figure 6: Cumulative Cases, as a Percentage of Population

A Boxplot
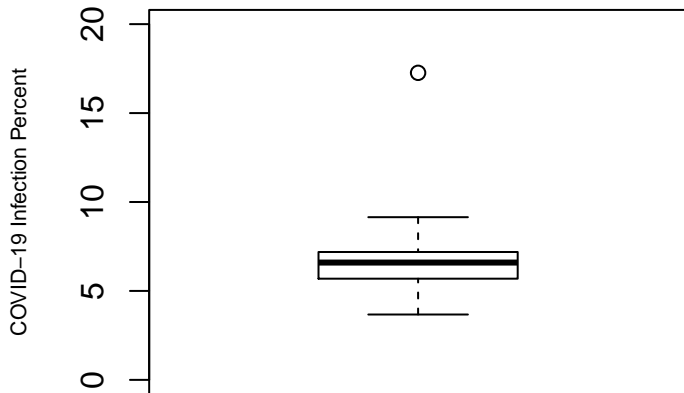
Figure 7: Cumulative Cases, as a Percentage of Population

Multiple Boxplots in One Figure
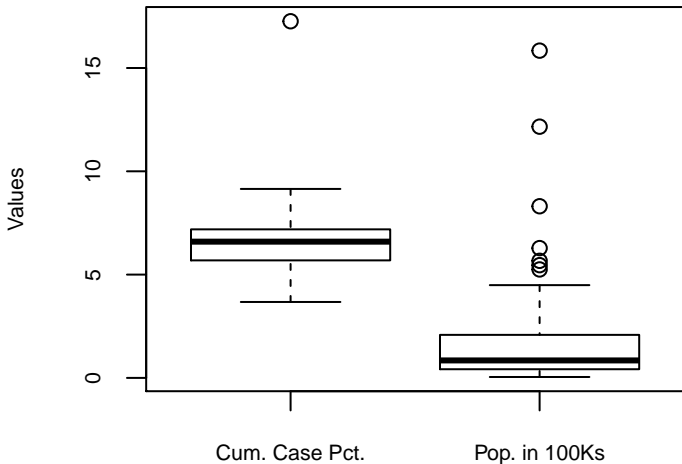
Figure 8: Two Boxplots

Quantile-Quantile (Q-Q) Plots

- Plots two sets of quantiles against each other...
- Can be used to compare two variables' distributions to each other
- Can also compare the *empirical* distribution of a variable to a *theoretical* distribution
- If the two are the same, the quantiles will lie on a straight line

One Density. . .

New COVID Cases per 100K
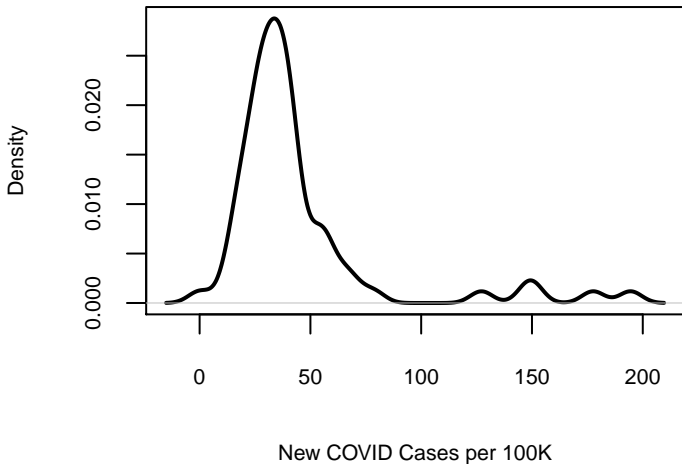
Figure 9: New Cases per 100K Population

One-Variable Normal Q-Q Plot
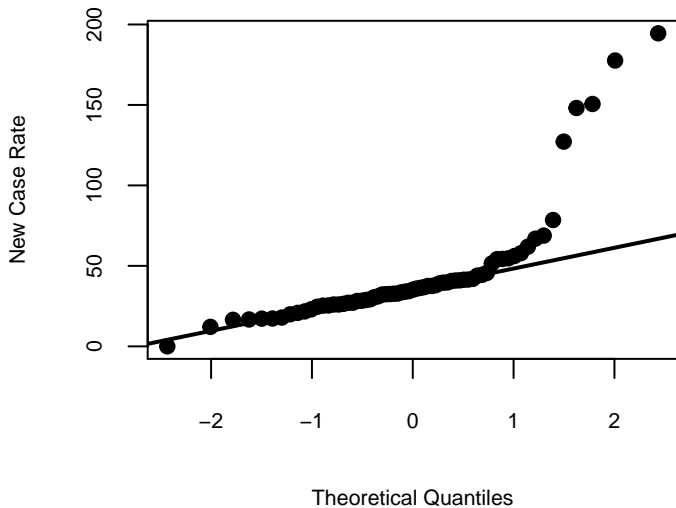
Figure 10: New Cases per 100K Population

Two Densities...
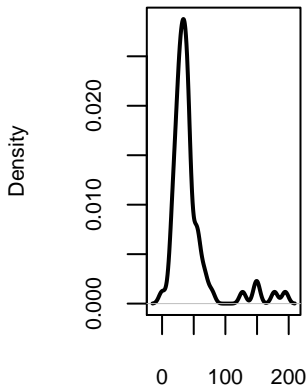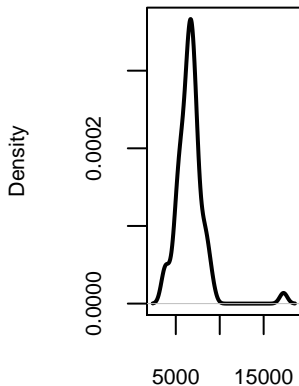
New COVID Cases per 100K   Cumulative COVID Cases per 100

Figure 11: New and Cumulative Cases per 100K Population
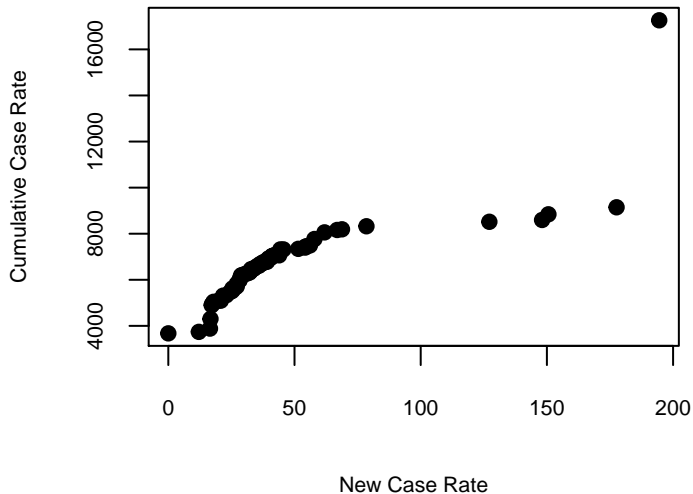
Two-Variable Q-Q Plot

Figure 12: New and Cumulative Cases per 100K Population

Other Univariate Plots

- Pie charts (please don't. . .)
- "Donut" plots (same)
- "Stem-and-leaf" plots (very old-school)
- Stripplots
- *Time-Series Plots*. . .

A Time Series Plot

Date
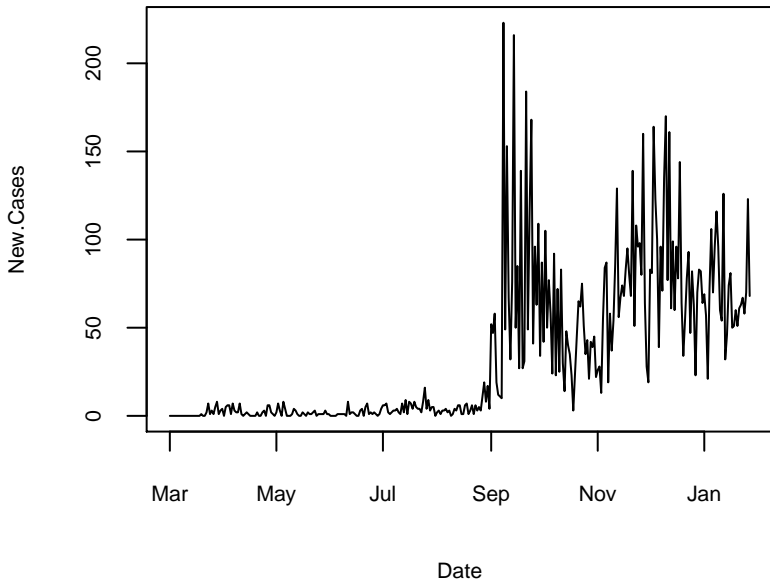
Figure 13: Daily COVID Cases in Centre County, 3/1/2020-1/27/2021

Add a "Smoother"

Date
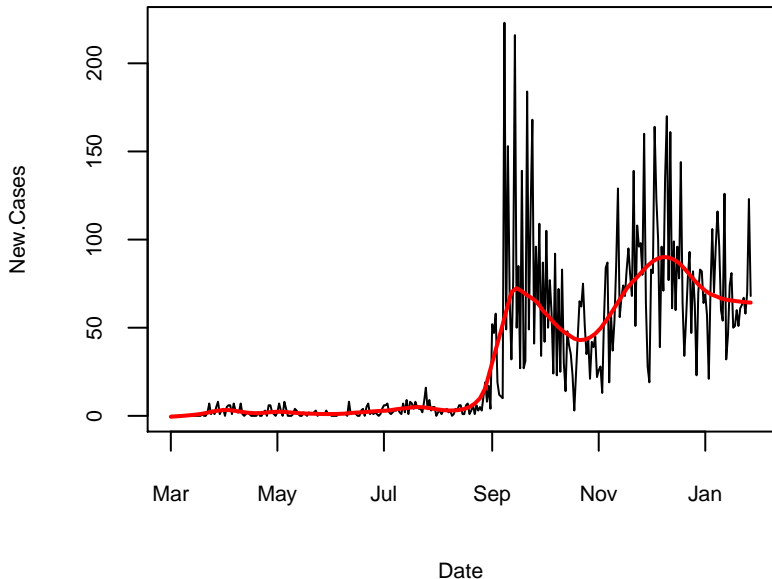
Figure 14: Daily COVID Cases in Centre County, 3/1/2020-1/27/2021
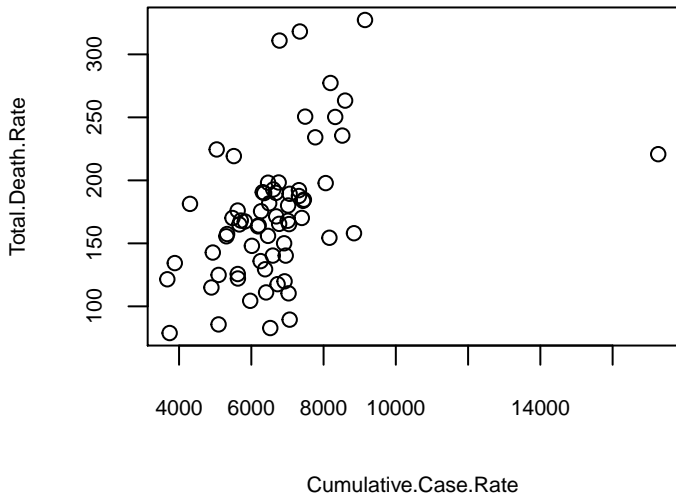
Bivariate Plots: The Scatterplot

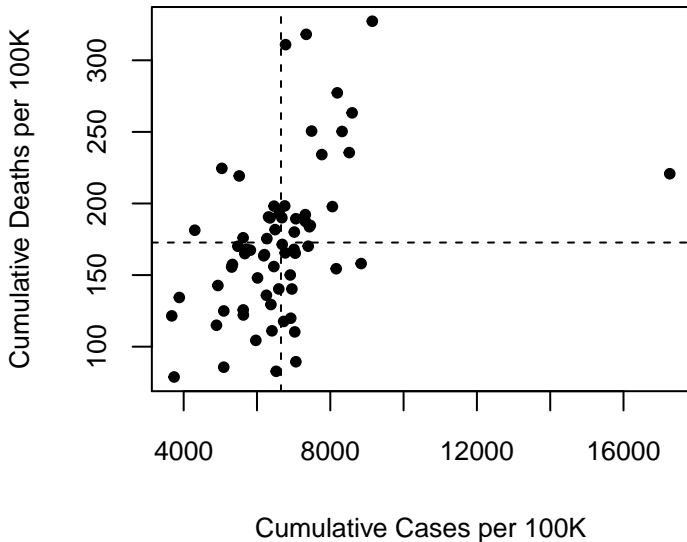Figure 15: Case Rates vs. Death Rates

A Better Scatterplot

Figure 16: Case Rates vs. Death Rates
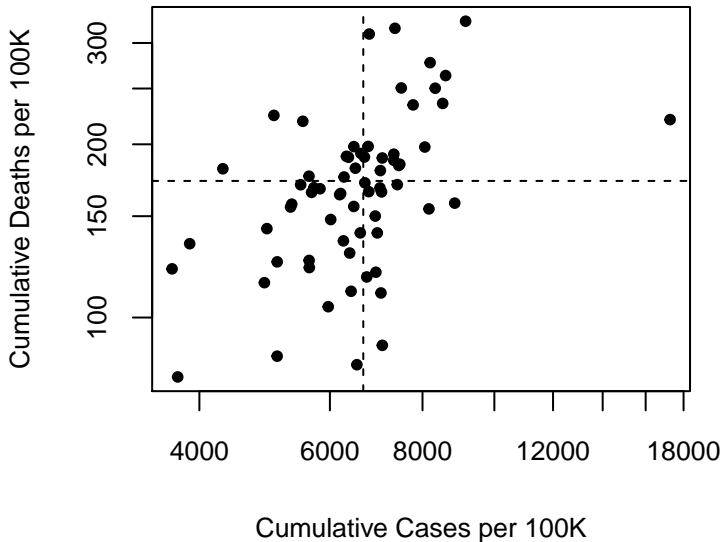
Rescaling Axes (Log Scales)

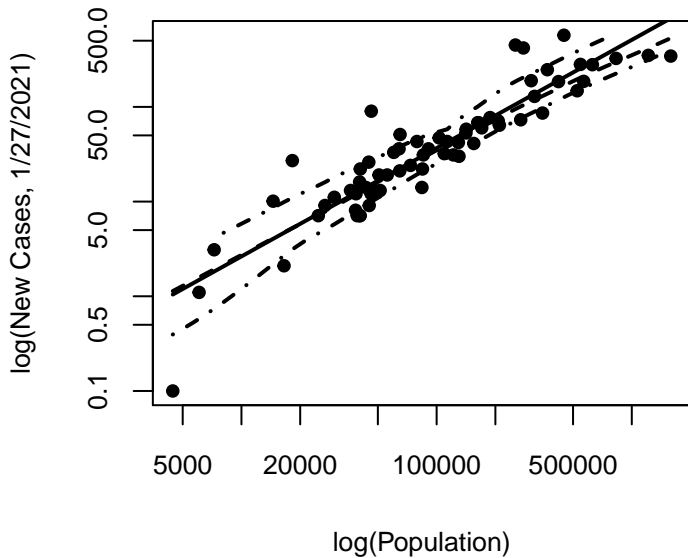Figure 17: Case Rates vs. Death Rates

Adding Lines

Figure 18: New Cases vs. Population
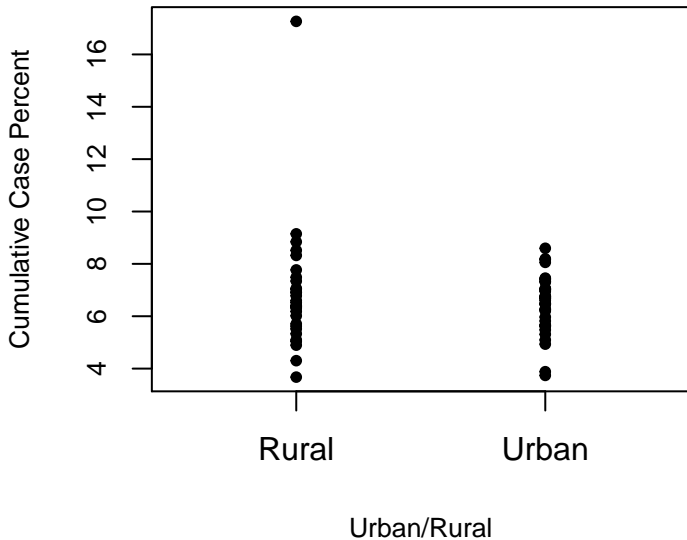
How Not To Draw A Scatterplot

Figure 19: Urban/Rural vs. Cumulative Case Percent
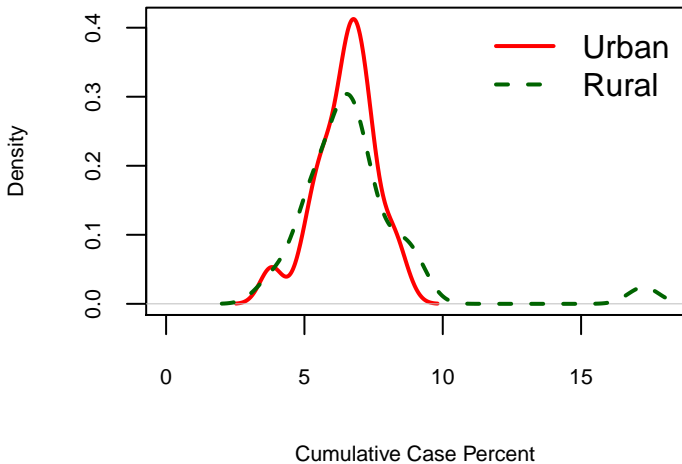
Better. . .

Figure 20: Cumulative Cases: Percent of the Population
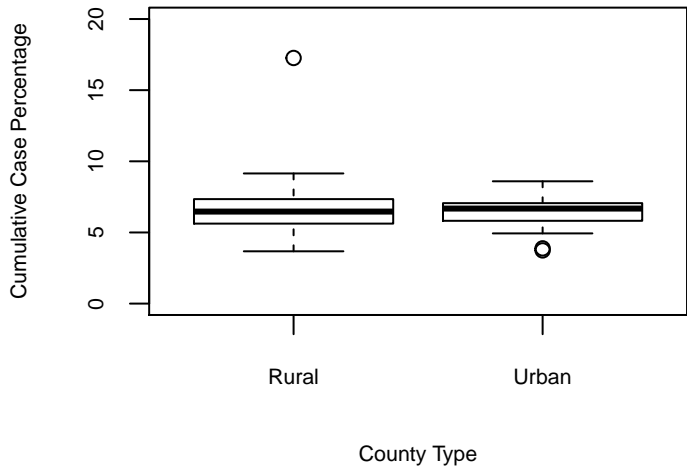
Or:

County Type

Figure 21: Cumulative Cases: Percent of the Population

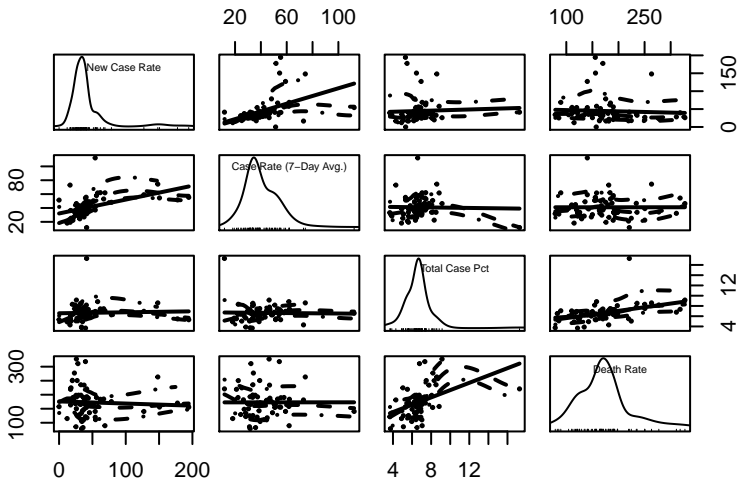## Multivariate Plots: Scatterplot Matrix
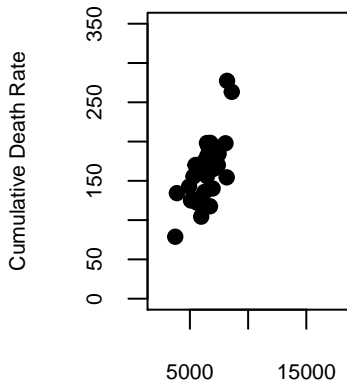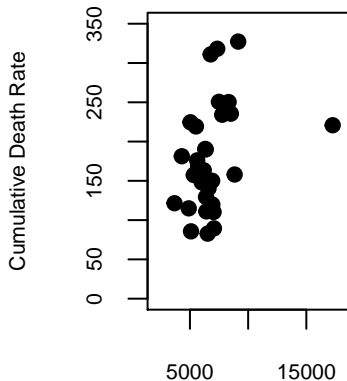

Figure 22: COVID in Pennsylvania (1/27/2021)

Conditioned Scatterplots

Cumulative Case Rate: Urban   Cumulative Case Rate: Rural

Figure 23: Case and Death Rates, by Urban/Rural

Other Cool Visualizations

- Contour / Wireframe Plots ("3D")
- Radar plots
- Parallel coordinates plots
- MAPS. . .
- Dynamic / interactive graphics
- Many more. . .