Data 1030 - Midterm Presentation

Laura McCallion, Brown University
October 25, 2019
https://github.com/lmccalli/1030_Project.git

Datasets:

- U.S. Census Bureau Quickfacts
- NNDSS - Table II. Varicella to West Nile virus disease
- NNDSS - Table II. Tetanus to Varicella
- Varicella Vaccination Coverage, 1996 to 2017
- Interactive map: https://www.cdc.gov/vaccines/imz-managers/coverage/childvaxview/data-reports/varicella/trend/index.html

Intro

Using Census data along with CDC monitoring programs to predict the rate of Varicella (chickenpox) cases.

Predicting outbreaks is critical. In addition, while the efficacy of vaccination doesn't need to be proved, the effects of vaccination rates should be monitored, especially given recent trends of vaccine-related misinformation.

Preprocessing

Data Assembly

- Required pulling together data sets from multiple sources
- Formatting, etc were different.
- CDC data in particular is valuable, but not user-friendly.

Missing Values

- The only missing values were in the target variable
- This is better than expected, but makes replacing them dubious
- Averages of each state were used

Types of Data

- All data was numeric
- MinMax Scaling was used for percentage- based data. Otherewise, standard scaling.

Not an overabundance of data points.

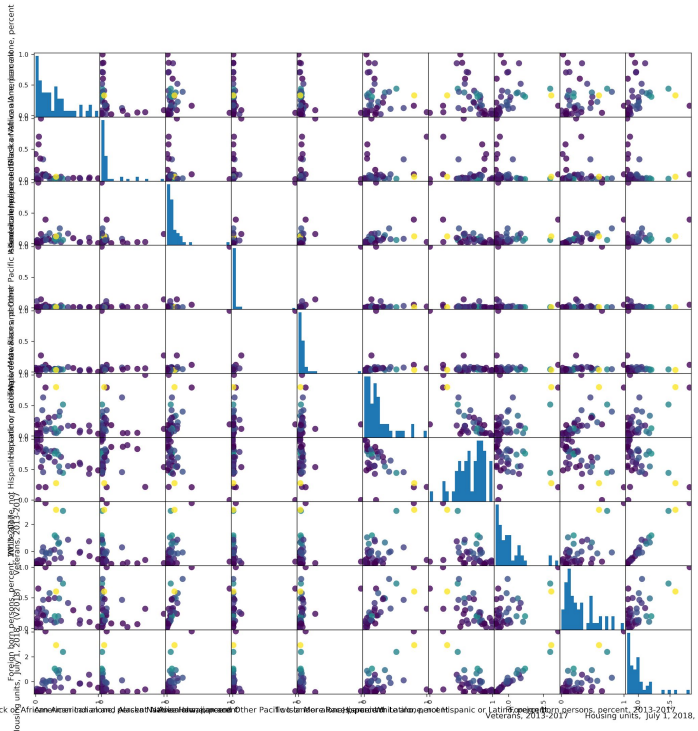
However, the case counts themselves are sizeable, which was part of the reason why I chose Varicella to analyze.

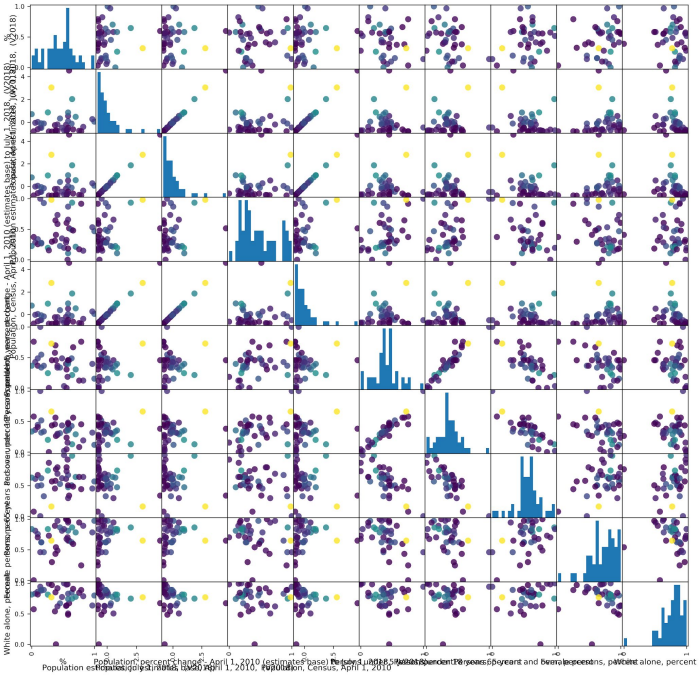
This is standardized CDC data, not census.

EDA - Basic Information

	Varicella (chickenpox), Cum 2017	Varicella (chickenpox), Cum 2016	Varicella (chickenpox), Cum 2014	Varicella (chickenpox), Cum 2013	Avg	n	%
count	36.000000	39.000000	38.000000	39.000000	39.000000	5.000000e+01	50.000000
mean	195.722222	229.076923	240.631579	291.000000	239.179487	3.247402e-16	0.456613
std	219.233289	261.837674	296.909798	348.305598	274.255667	1.010153e+00	0.232399
min	8.000000	5.000000	4.000000	5.000000	6.500000	-1.097654e+00	0.000000
25%	37.500000	47.500000	47.500000	57.500000	47.500000	-5.566956e-01	0.316532
50%	109.000000	149.000000	171.500000	168.000000	157.500000	-3.230998e-01	0.475806
75%	279.250000	309.500000	309.750000	409.000000	343.375000	1.287236e-01	0.580645
max	953.000000	1341.000000	1557.000000	1874.000000	1431.250000	3.242309e+00	1.000000

EDA - A Broad View

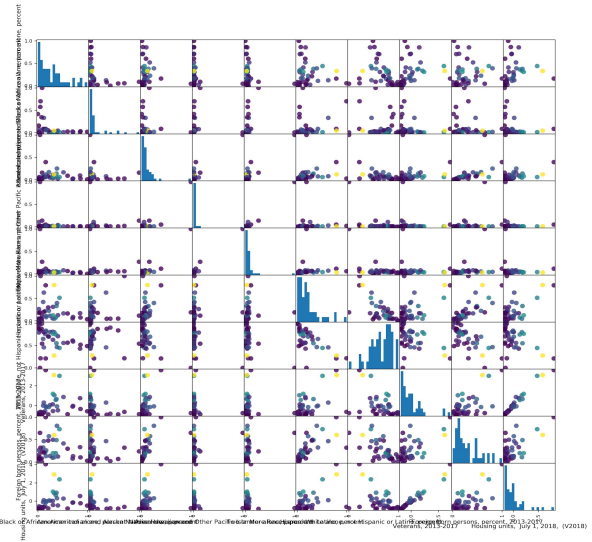


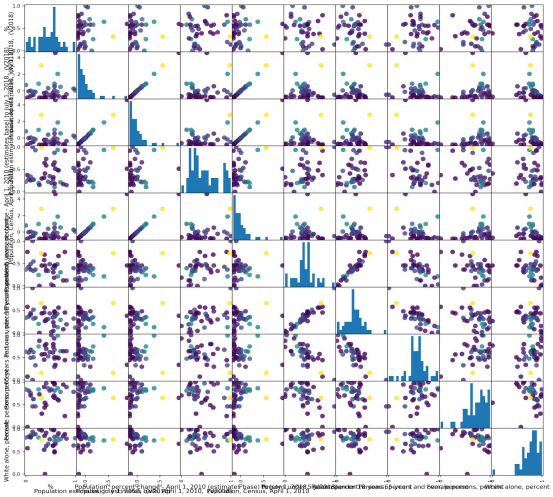


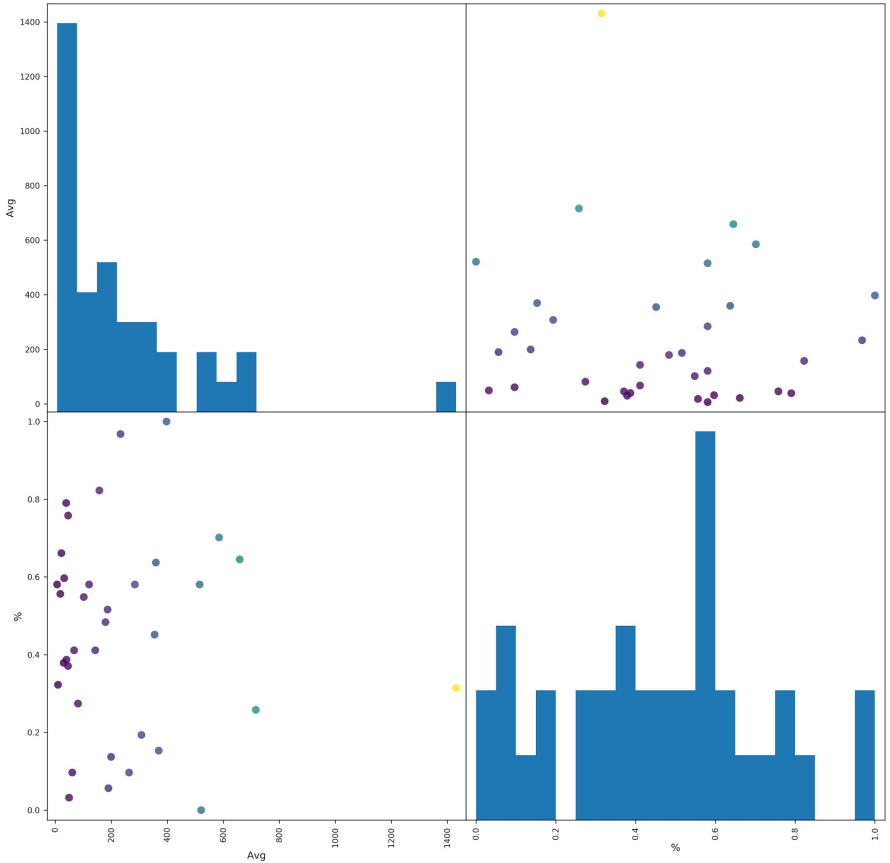
EDA - A Broad View

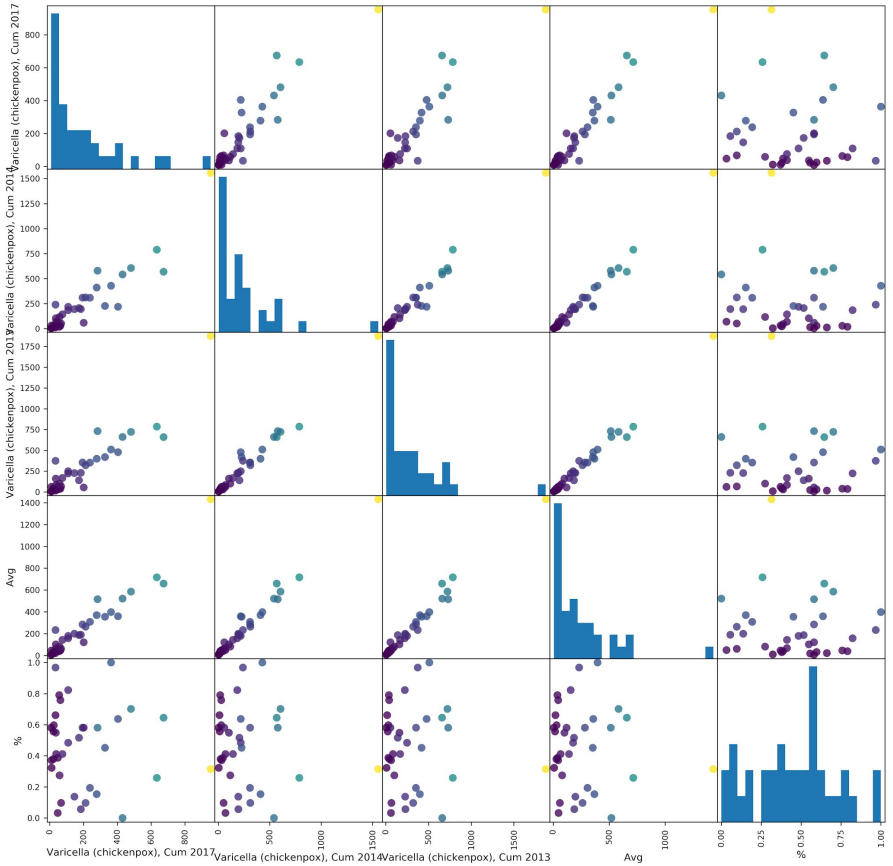
Notable trends:

- Color is based on target variable, and most of the charts show some sort of trend
- A lot of the scatter plots seem to show a high level of correlation between variables (we will see this later as well)







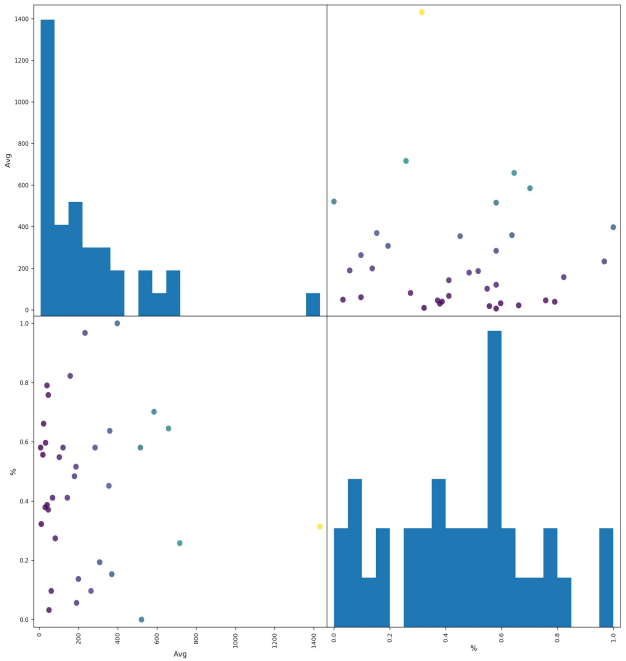


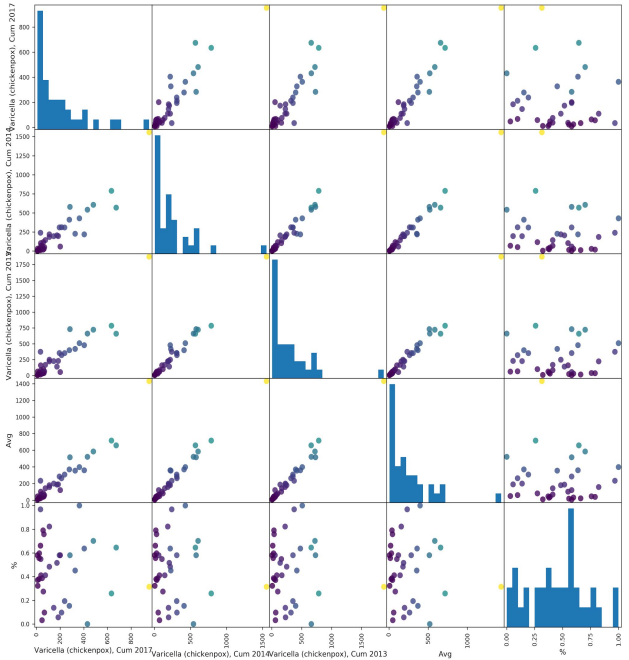
EDA - The Target Variable

EDA - The Target Variable

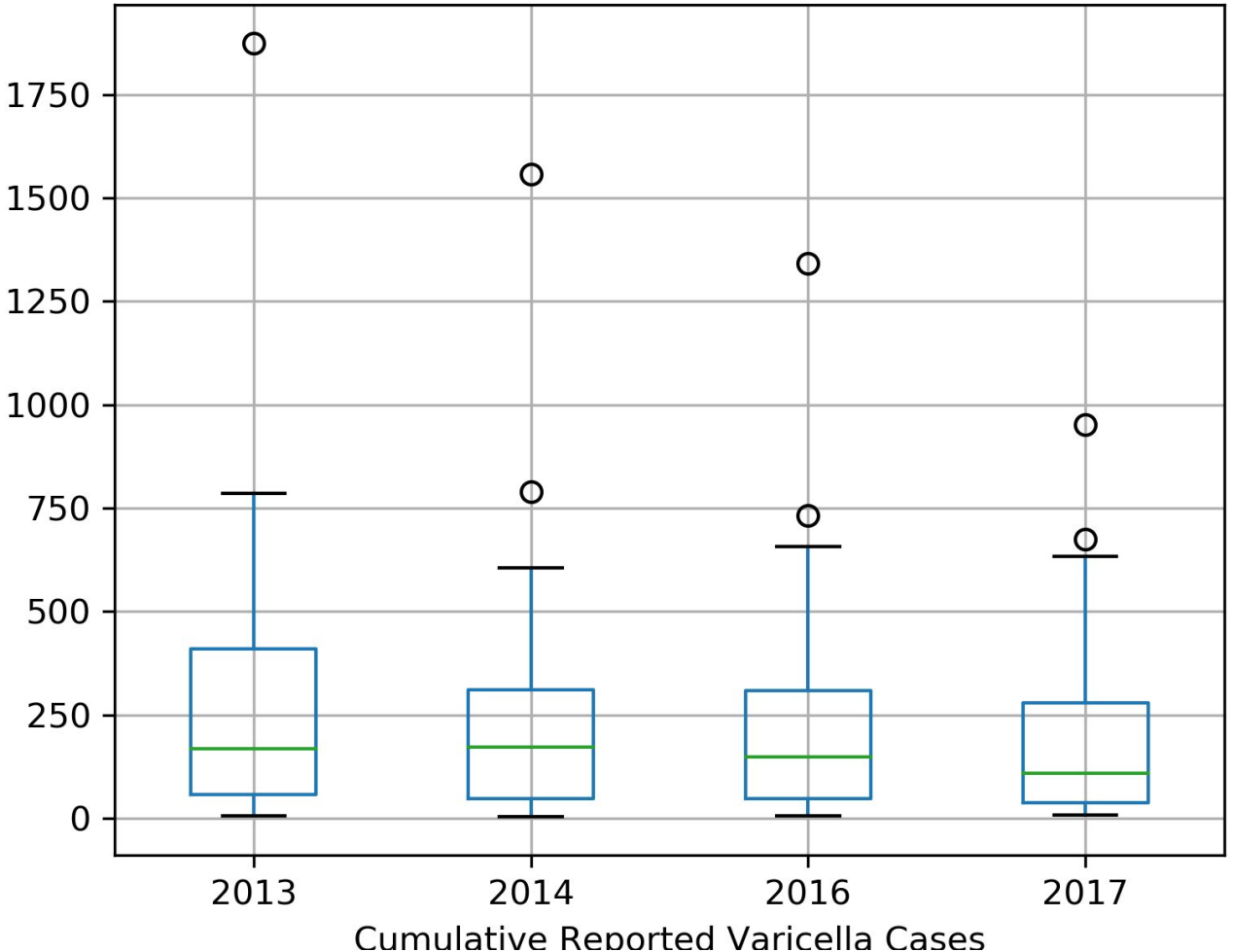
On the left, you can see the average cases compared to the scaled vaccination coverage. The pattern isn't as distinct as might be expected, but still shows that the big outbreaks don't tend to occur in highly vaccinated areas.

The right compares the counts for each year to each other. You can see that these relationships appear near linear.



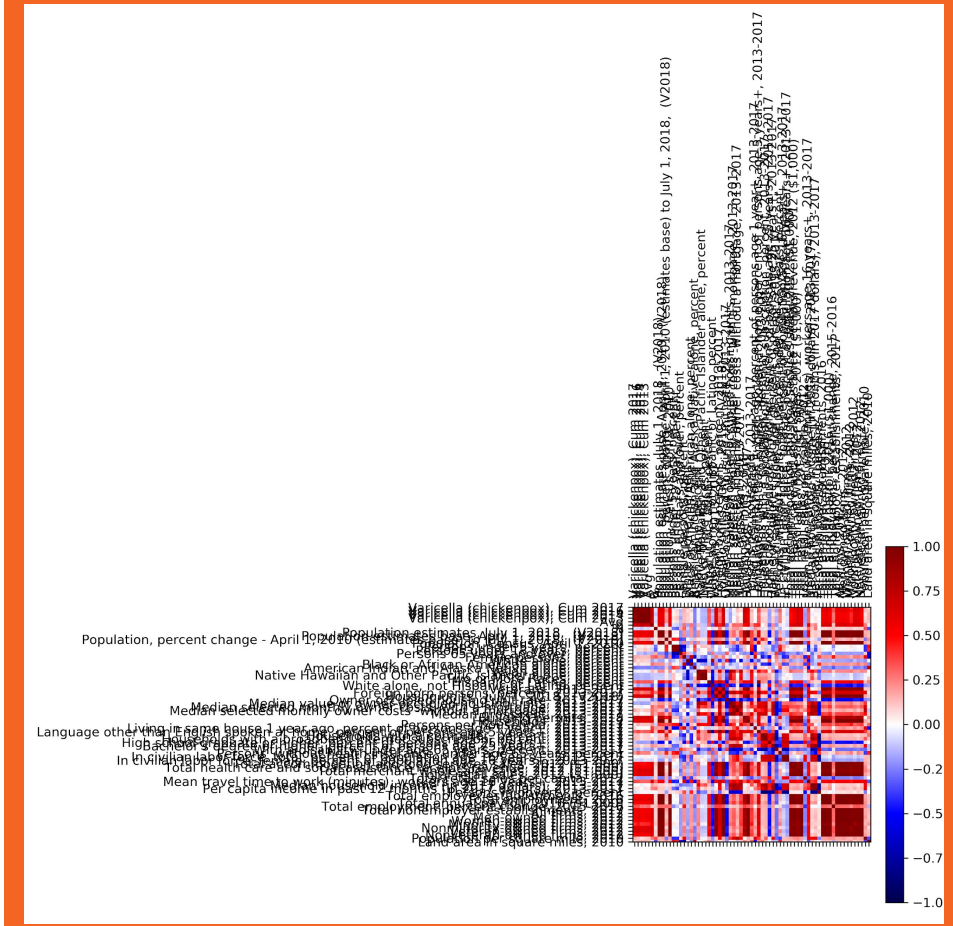


Here, we see that the reported cases seems to decrease over time, which is good news. However, decrease is small, and the means remain close.



Correlations

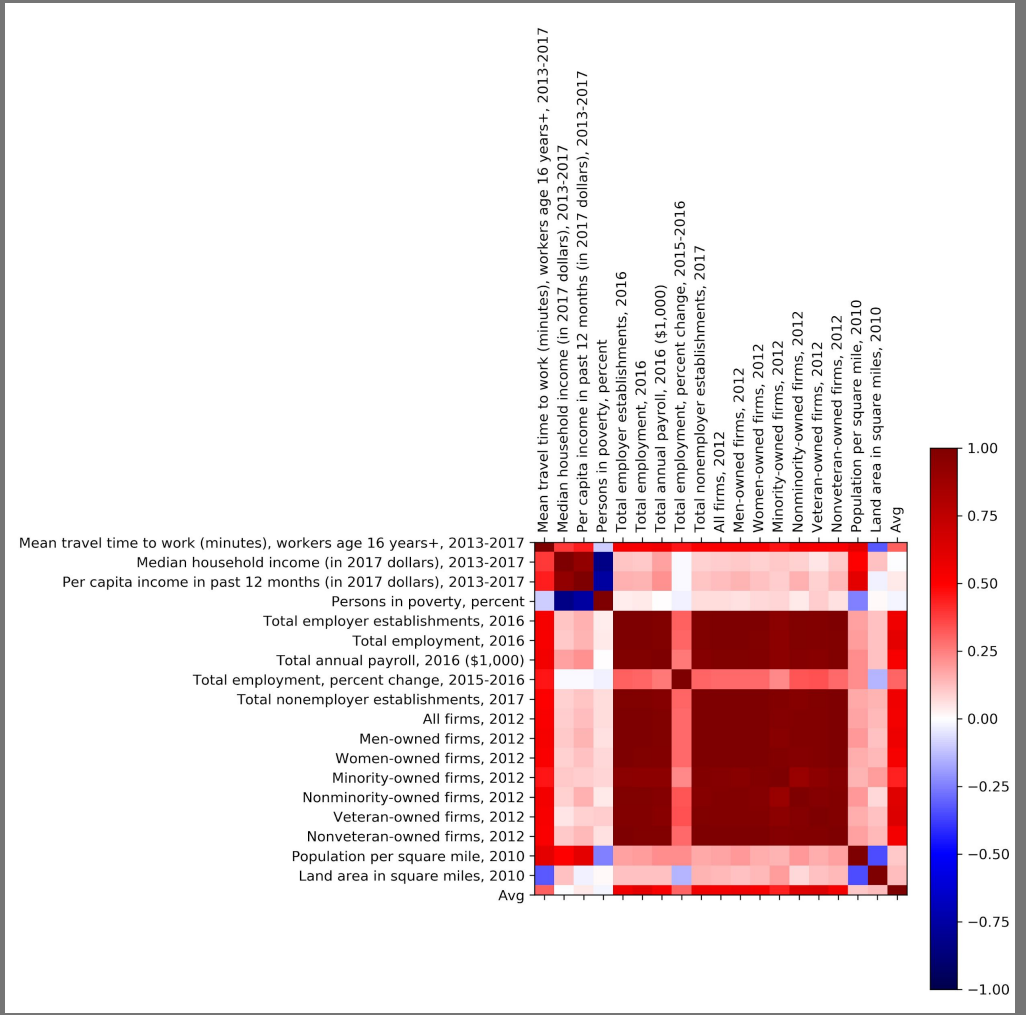
While this matrix is too large to make out any specific trends, you can see patterns of highly correlated areas.



—

Zooming in, wee can see that the correlation comees from the census data.

The census bureau collects a lot of data on very similar areas, and these areas are imported together in the dataset. Hence, the large blocks of highly correlated variables.



—

This might end up being a problem, however.

This chart shows the most correlated variables to the target of average reported cases.

A lot of these variables are extremely similar to each other.