Supervised Learning Unit 3 Capstone

By Grace Flynn

Purpose And Target

- If we examine a city's health behaviors and disease proportion, can we estimate the percentage of people with/without health insurance?

- Target Variable = Percent of People with Access to Healthcare

Terms

- Comorbidity – when two or more illnesses are found together
  - Depression and anxiety are a common example of this
- Access – access to healthcare can be measured by a number of things, but this analysis defines it as not having health insurance
- PCP – Primary care physician

500 Cities CDC Dataset – What's in it?

- Measures:
  - Unhealthy behaviors
  - Health Outcomes
  - Items that could prevent health problems
- How
  - Uses information from CDC Behavioral Risk Factor Surveillance System, the 2010 census, and American Community Survey estimates
  - Estimates are used from a combination of multi-level regression & poststratification (MRP), geocoded health surveys, and socioeconomic data.
  - www.cdc.gov/500cities/methodology

Exploratory Data Analysis

-Three different measurements for each variable:

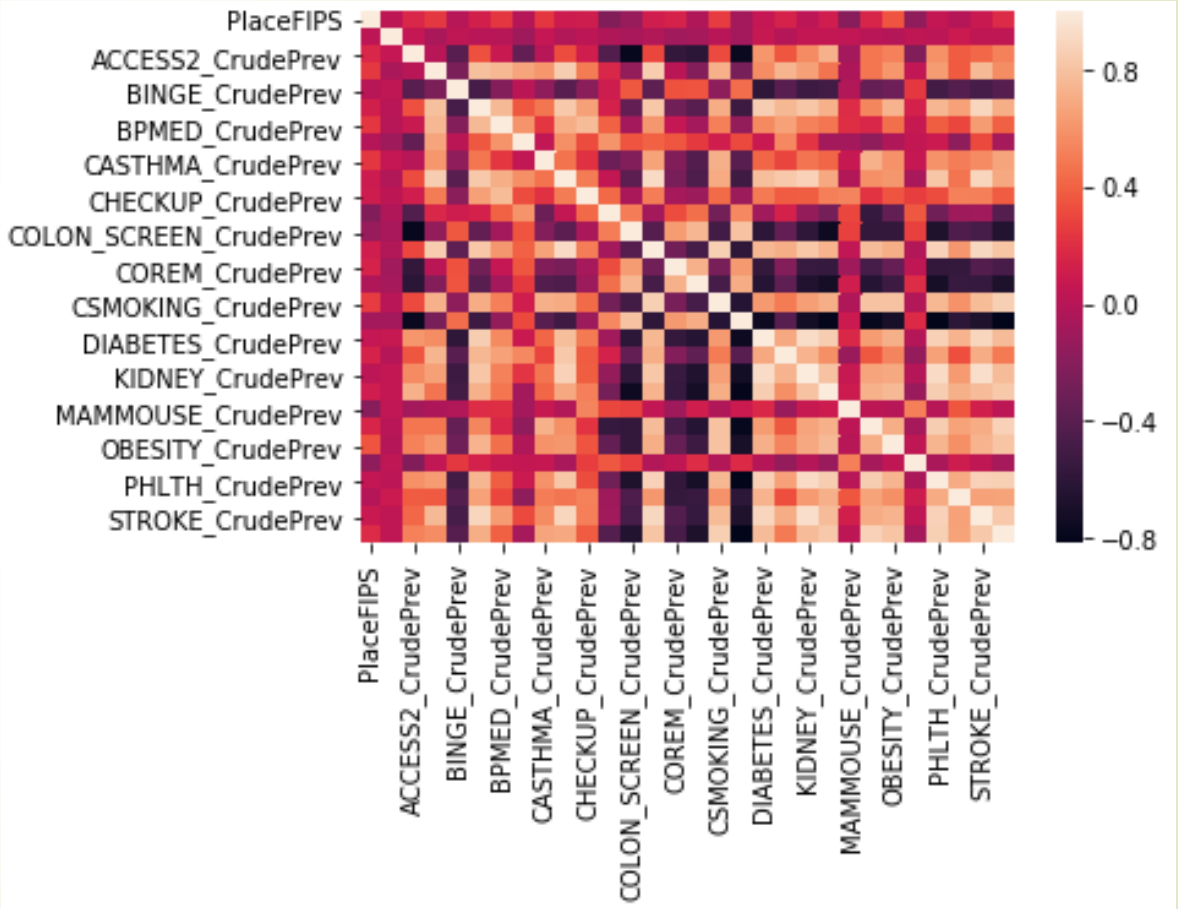
1. Crude
2. A 95% confidence interval
3. Adjusted

	StateAbbr	PlaceName	PlaceFIPS	Population2010	ACCESS2_CrudePrev	ACCESS2_Crude95CI	ACCESS2_AdjPrev	ACCESS2_Adj95CI	ARTHRITIS_CrudePrev	ARTHRITIS_Crude95CI	ARTHRITIS_AdjPrev	ARTHRITIS_Adj95CI
0	CA	Folsom	624638	72203	7.5	(7.0, 8.0)	7.7	(7.2, 8.2)	16.9	(16.6, 17.2)	17.4	(17.2, 17.7)
1	FL	Largo	1239425	77648	19.6	(19.1, 20.2)	20.9	(20.4, 21.5)	30.6	(30.3, 30.9)	23.3	(23.1, 23.5)
2	CA	Berkeley	606000	112580	7.7	(7.3, 8.1)	7.1	(6.8, 7.3)	15.1	(15.0, 15.3)	18.0	(17.8, 18.1)
3	CA	Napa	650258	76915	12.3	(11.8, 12.8)	12.7	(12.1, 13.3)	20.7	(20.5, 21.0)	19.3	(19.1, 19.5)
4	FL	Sunrise	1269700	84439	22.8	(22.1, 23.5)	23.3	(22.6, 24.1)	22.8	(22.5, 23.1)	20.8	(20.6, 21.0)

5 rows × 117 columns

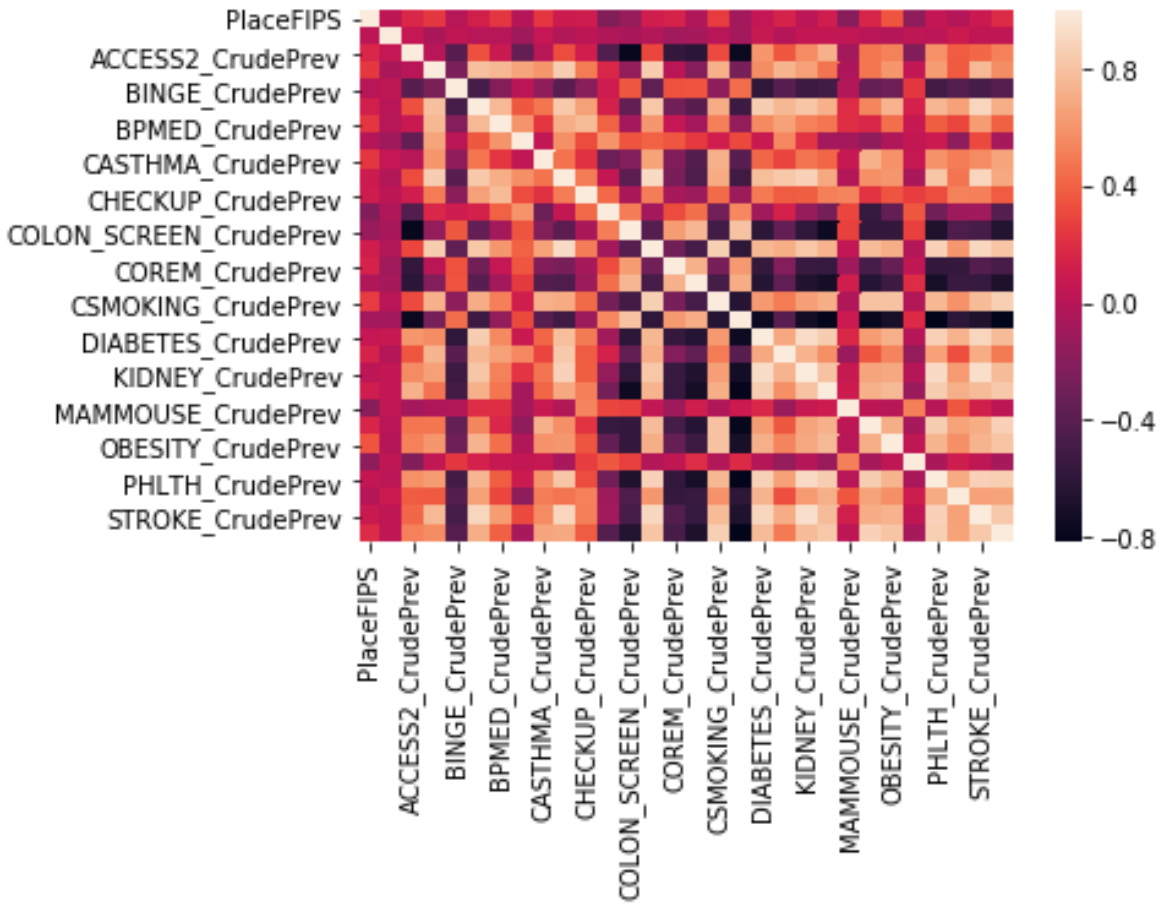
Correlation check

- Highly correlated items
- Need to be dropped to avoid multicollinearity
- Method used: anything >=.6 dropped

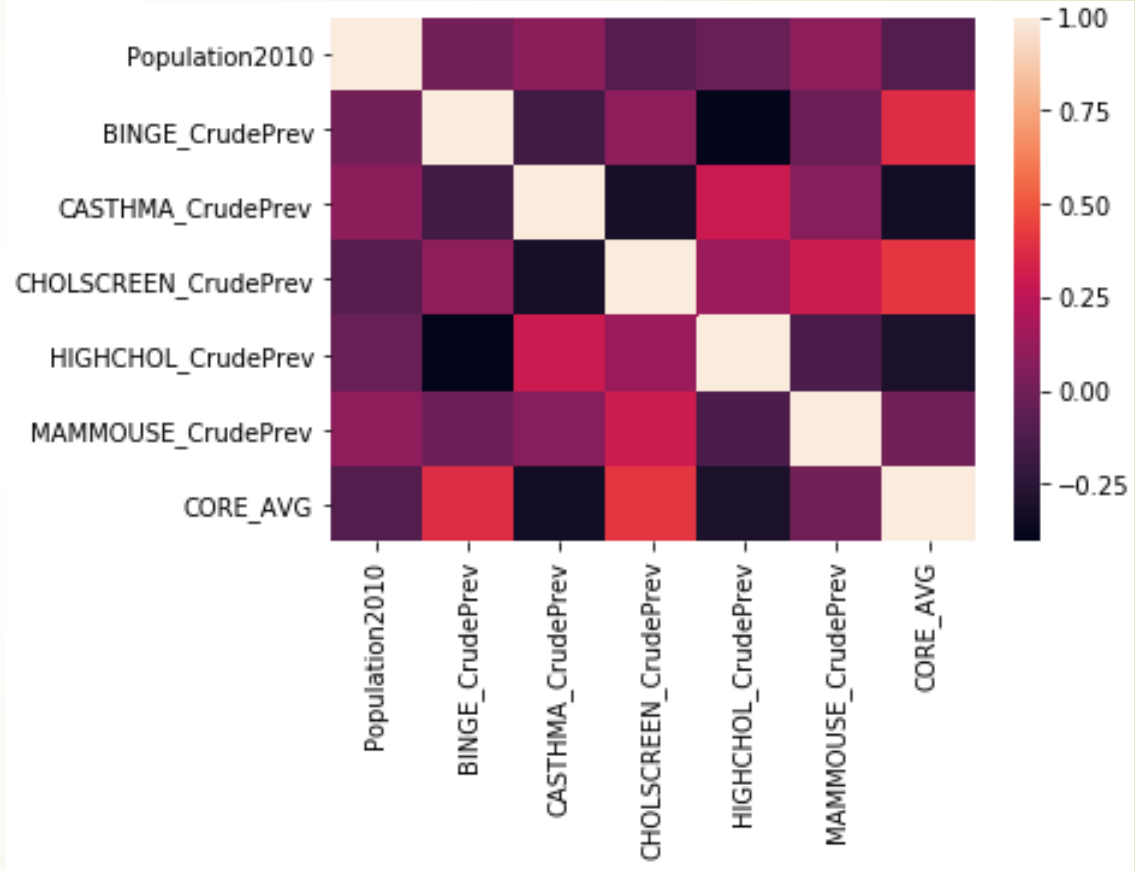


(df.drop it like it's hot)

Before dropping

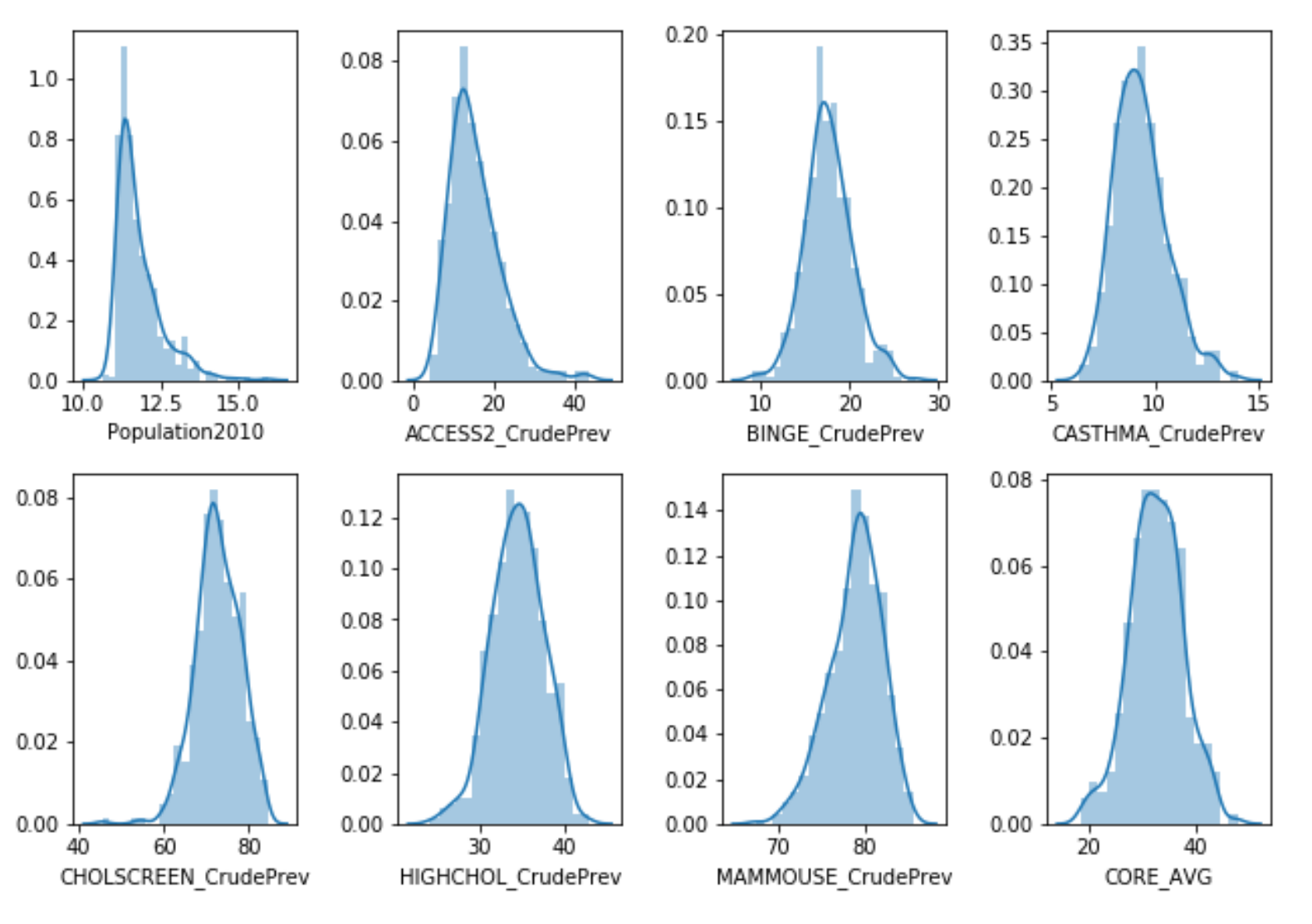


After dropping



Hot like a heat map

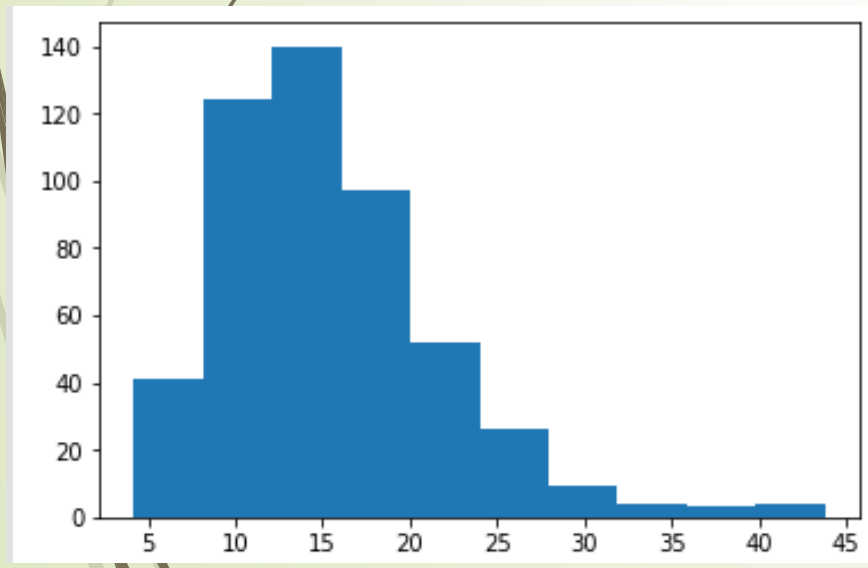
7 Features

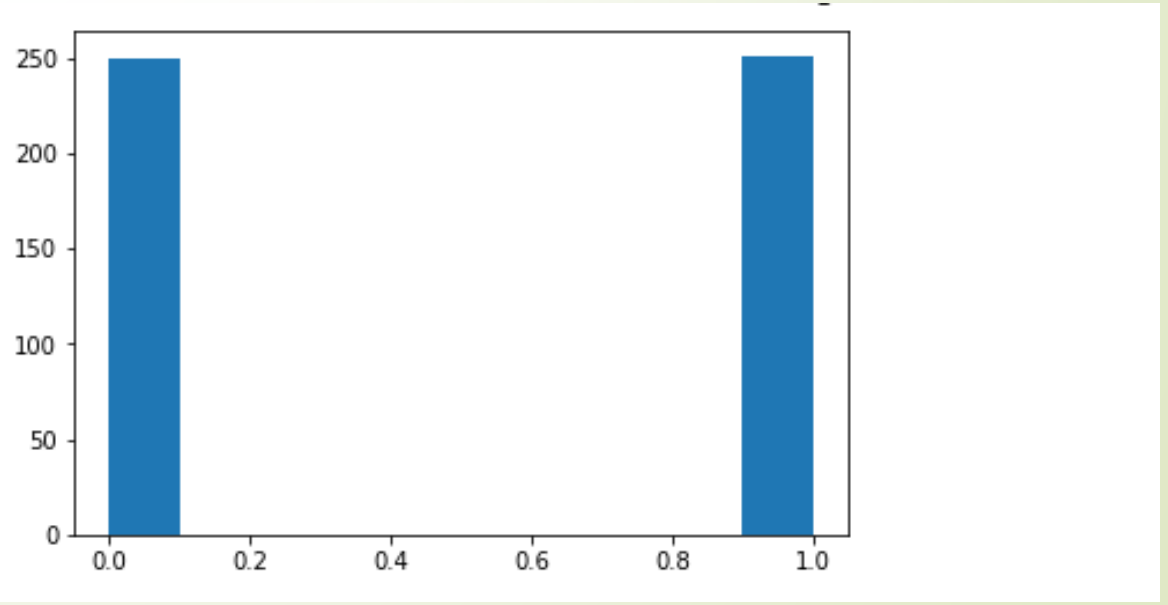

Population2010
ACCESS2_CrudePrev
BINGE_CrudePrev
CASTHMA_CrudePrev
CHOLSCREEN_CrudePrev
HIGHCHOL_CrudePrev
MAMMOUSE_CrudePrev
CORE_AVG

Dependent Variable Transformation

count	500.000000
mean	15.453800
std	6.268764
min	4.200000
25%	11.100000
50%	14.100000
75%	18.700000
max	43.800000

- Below the median = low-risk
- Above the median = hi-risk





Models and Results

Test Used	CV = 3 Result
Logistic Regression	[0.79041916, 0.83233533, 0.76506024]
Decision Tree	[0.66467066, 0.67065868, 0.6746988]
Random Forest	[0.73053892, 0.77245509, 0.76506024]
Random Forest with Grid Search CV	[0.70658683, 0.76047904, 0.70481928]
Gradient Boosting	[0.70658683, 0.74251497, 0.71686747]

Best Model for classification by r^2

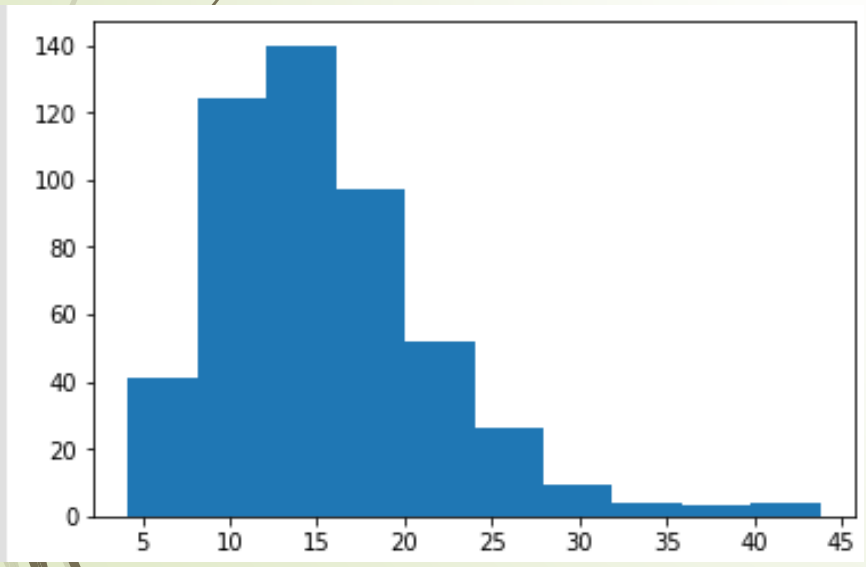
Test	Train Score	Test Score	CV=3
Logistic Regression	0.806	0.827	[0.79041916, 0.83233533, 0.76506024]

Confusion Matrix score:

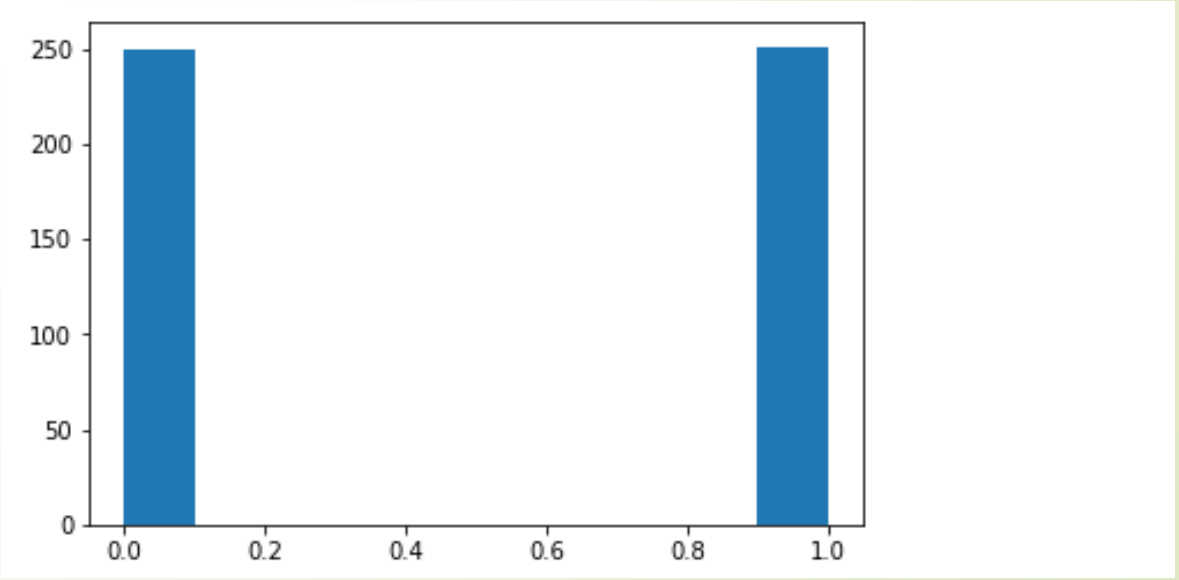
	1	0
1	61	10
0	20	59

Redefine dependent variable...

count	500.000000
mean	15.453800
std	6.268764
min	4.200000
25%	11.100000
50%	14.100000
75%	18.700000
max	43.800000

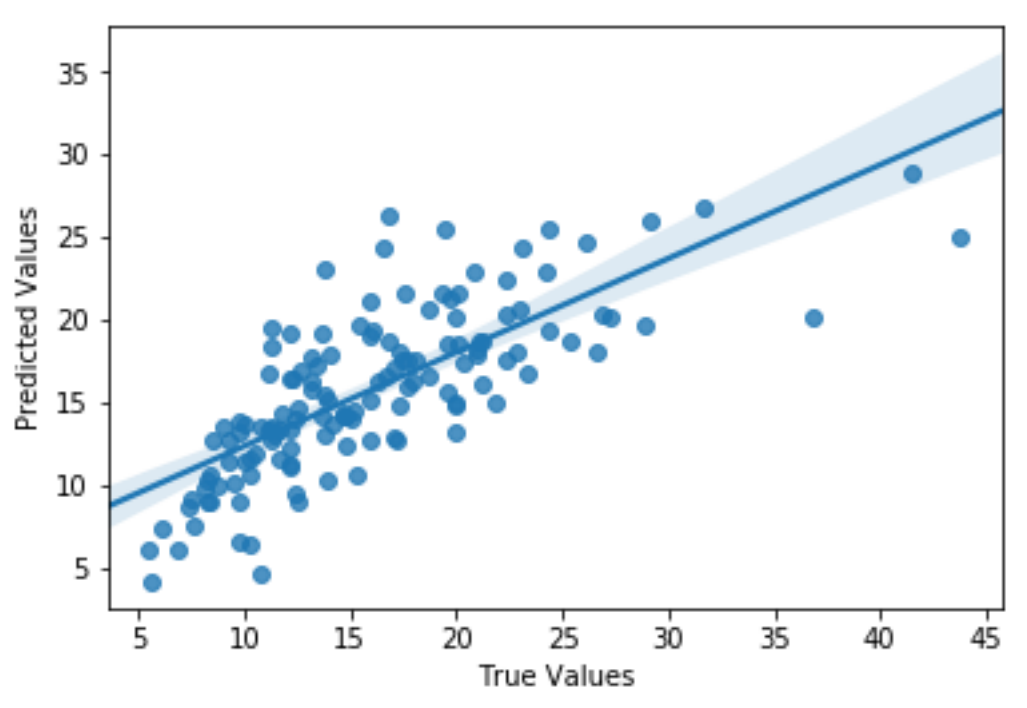


- Below the median = low-risk
- Above the median = hi-risk



Leave as a continuous variable

- Ran PCA (n_components = .80) and then vanilla linear regression:



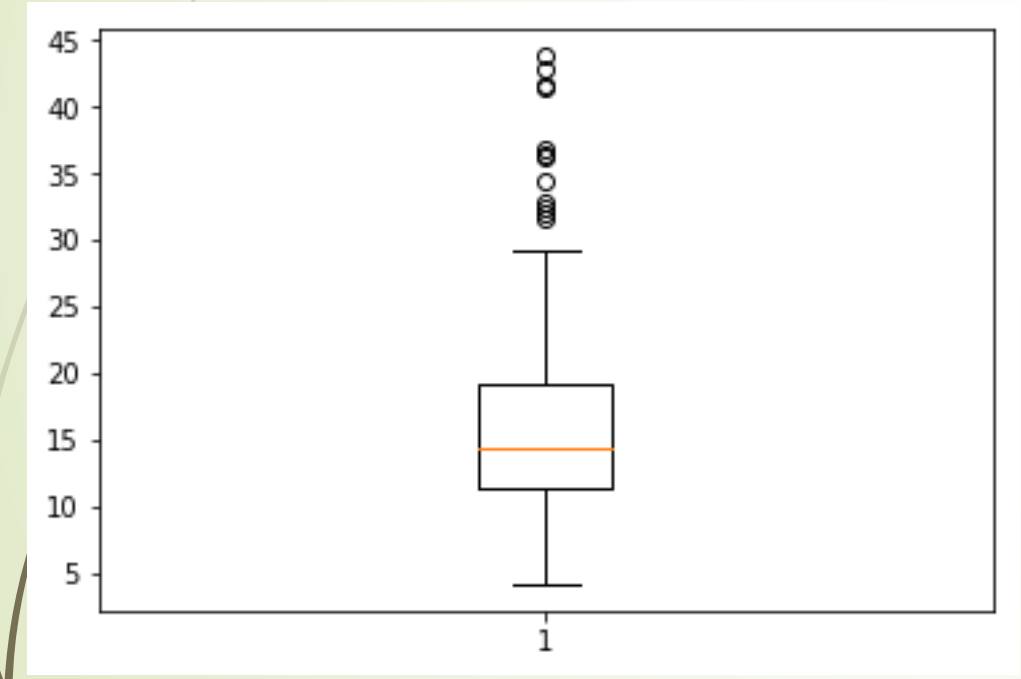
Train Score	Test Score	CV = 3
0.625	0.572	[0.59025356, 0.61769295, 0.58687229]

Could be improved!

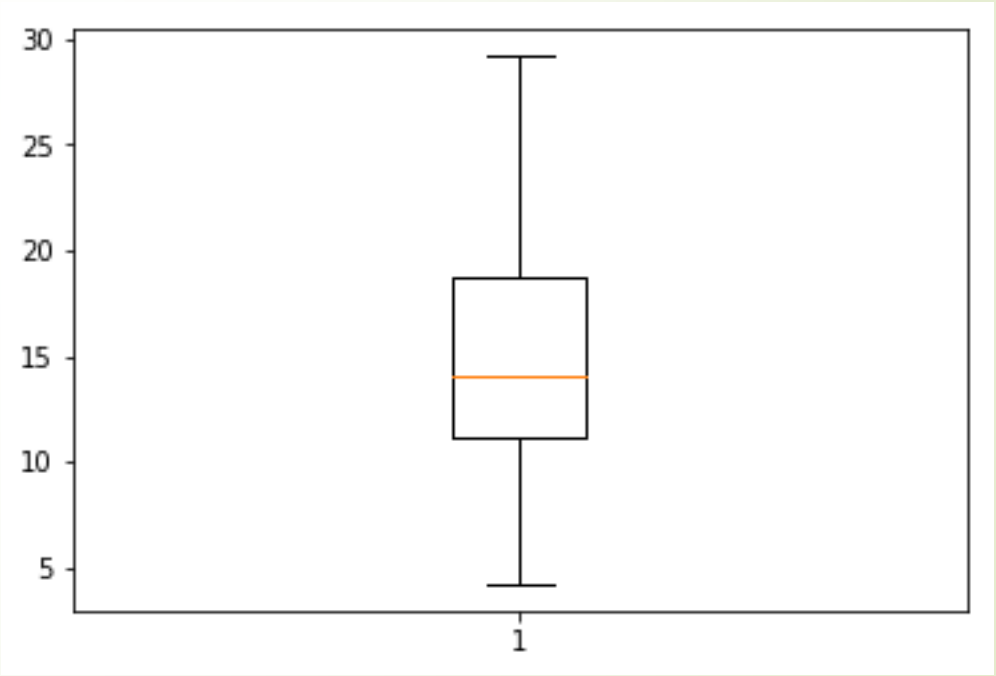
Dealing with outliers…

n outliers = 12

Before:



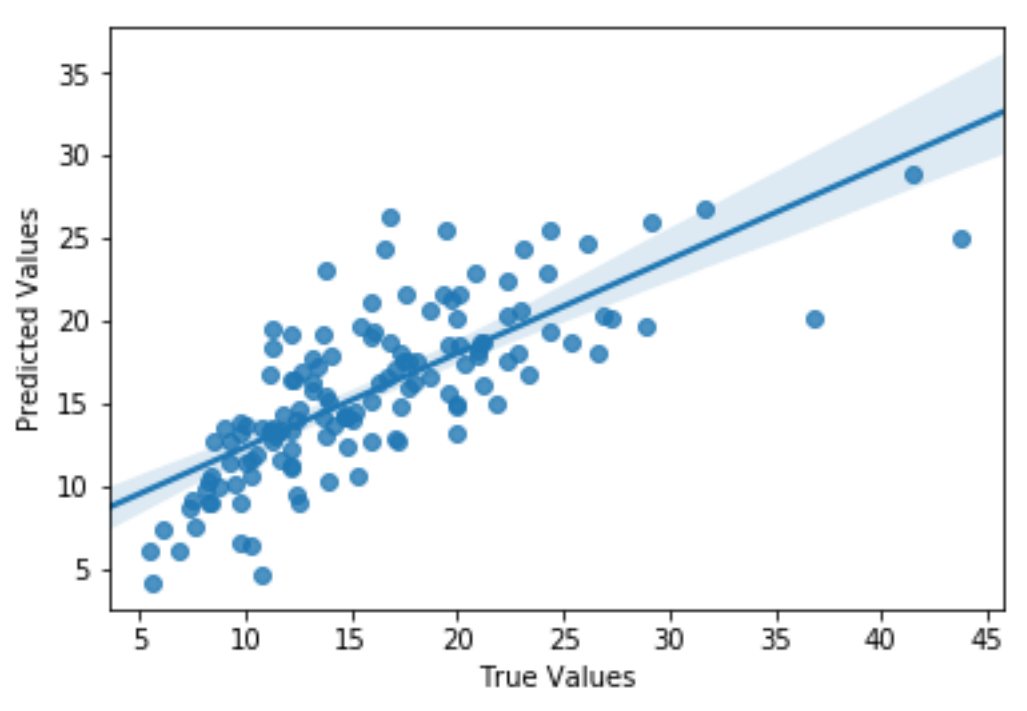
After:



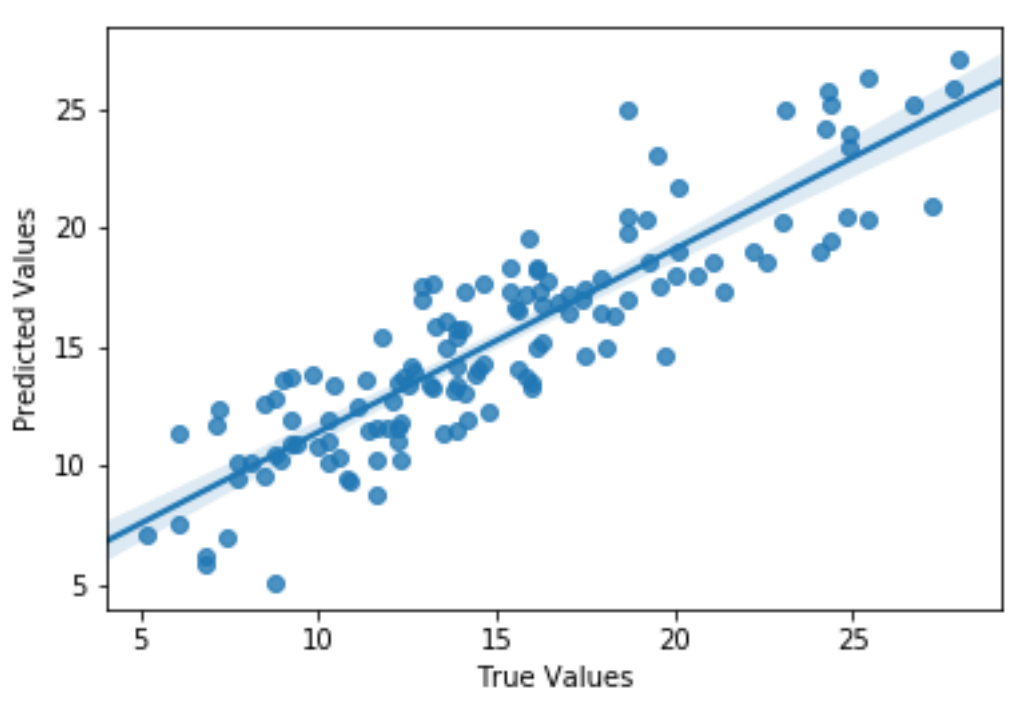
Dealing with outliers…

n outliers = 12

Before:



After:



Last Table

	Train Score	Test Score	CV = 3
With outliers	0.625	0.572	[0.59025356, 0.61769295, 0.58687229]
Without outliers	0.732	0.784	[0.74706727, 0.73632426, 0.74497987]

Assumptions and Weaknesses

- CDC made accurate predictions from survey and census data
  - Poststratification method
- CDC accurately corrected for non-response bias
  - Those severely in poverty and/or afflicted with multiple illnesses are probably more likely to not respond to surveys
- The CDC chose 500 cities that accurately represent the US
- Lost data in making this a classification problem
- Linear regression lost data by removing outliers, but increased accuracy

Further Research

- Acquire more data and analyze as time series
- Compare data across countries with different healthcare systems