**Predicting Local Experience after the Great Recession**

Kyle Alden -- DAT8

**1. Background**

Localities experienced the Great Recession in vastly different ways. Some economies quickly returned to their pre-recession health, while others have yet to fully regain their footing 6 years after the official end of the recession.

As of 2014 only 15% of counties in the United States had returned their pre-recession unemployment rate, while only 56% percent had returned to within 1% of pre-recession unemployment. Over 30% of counties still saw unemployment over 2% higher than the average rate for 2007.

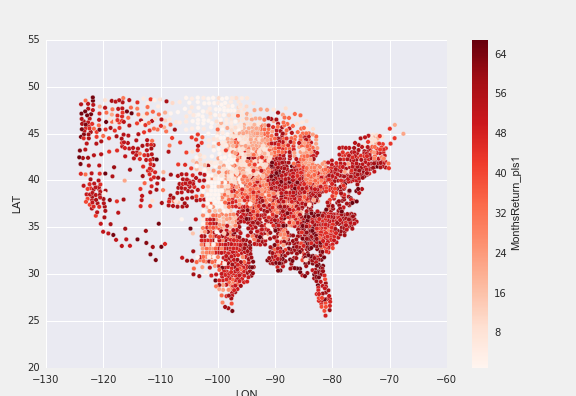
**2. Problem Statement**

Can I predict whether any county in the US to return to within 1% of its pre-recession unemployment rate (average rate in 2007)? Also, for counties that have returned can we predict how many months it will take?i.e. What types of counties are successful in America’s post recession economy?

2.1 Initial Hypotheses:

1. Rural, suburban and urban counties will experience the recession differently
2. Local industry matters: counties that depend on farming and energy production probably fared better than those dependant on the service and government sectors.
3. Seasonality probably matters

**Months until return within 1% of pre-recession unemployment**



**3. Response Variable Development**

**3.1 Data Source**

Data was provided as a flat file of monthly unemployment rates from the Bureau of Labor statistics. The file was comprised of a series\_id (every geographic area, with a code for data type), year, period(month), footnotes, and value--estimated unemployment rate, labor force, or number of unemployed individuals. Original data is available at <http://download.bls.gov/pub/time.series/la/>. The “la.data.64.County” file at the link contains the entire historical series for all counties it was not included in my github repository because of the 100mb limit.

**3.2 Regression Response Variable Engineering**

The original data had a shape of (4272544, 6). We only needed data from after 2007 and that was at the county level. This was accomplished using boolean operations in a pandas dataframe where we selected out all features that were from after 2006 and their series\_id ended with code “3”. This left a dataframe with a shape of (357231, 6).

Next, in order to create a dataframe where I could easily calculate the average 2007 unemployment rate and when a county returned to that rate, I needed a dataframe where each row represents a county and each column represents the unemployment rate of that county during that month. I calculated a YearMonth value that would be the by concatenating in pandas. I had some issues creating a pivot table in pandas, so to expedite the process I saved the data as a csv and used google sheets to further develop my variables.

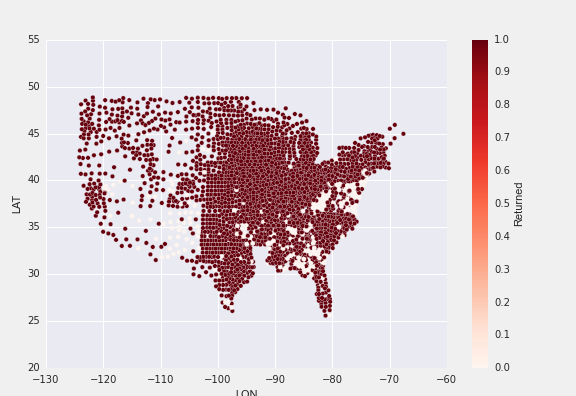
In Sheets, I created our pivot table, found the average unemployment rate for 2007 for each county by using =average(). I then calculated the month when a county’s unemployment rate dipped below the average using, =INDEX(AM$1:DN$1,MATCH(TRUE,INDEX(AM4:DN4<C4),0)) to get the column name (YearMonth) of each month. I found several anomalies where a county’s unemployment rate would drop and then jump back up, so I chose to repeat the process on a 3 month moving average of the rate.

Using the result of the INDEX operation--the column name representing the date--I then calculated the number of months past 1 January 2010 (the beginning of the recovery for the purposes of this study). This was then coded as ‘MonthsReturn\_pls1’.

**3.3 Classification Problem Response Variable Engineering**

I used the Months to return to pre-recession unemployment variable to calculate a binary variable of if a county returned to within one percent of its pre-recession uneployment rate or not. We gave a value of 1 to the counties that recovered and 0 to those that didn’t.

**Counties that returned to within 1% of 2007 unemployment rate**



**4. Feature Variables: Engineering and Exploration**

**4.1 Feature Joining Process**

The US Government uses a five-digit Federal Information Processing Standard (FIPS) code (FIPS 6-4) which uniquely identifies counties and county equivalents in the United State. These codes were available in our Series\_id variable downloaded from the Bureau of Labor Statistics. In order to join each of the following variables to our dataset I used the VLOOKUP operation in google sheets. Using a GUI for this data helped me quickly check my data.

**4.1.1 Demographic / Socio-economic Features**

These variables were downloaded from the US Department of Agriculture

at <http://www.ers.usda.gov/data-products/county-level-data-sets/download-data.aspx>

|  |  |
| --- | --- |
| **Rural\_urban\_continuum\_code\_2013** | The 2013 Rural-Urban Continuum Codes form a classification scheme that distinguishes metropolitan counties by the population size of their metro area, and nonmetropolitan counties by degree of urbanization and adjacency to a metro area. The official Office of Management and Budget (OMB) metro and nonmetro categories have been subdivided into three metro and six nonmetro categories. Each county in the U.S. is assigned one of the 9 codes. This scheme allows researchers to break county data into finer residential groups, beyond metro and nonmetro, particularly for the analysis of trends in nonmetro areas that are related to population density and metro influence. The Rural-Urban Continuum Codes were originally developed in 1974. They have been updated each decennial since (1983, 1993, 2003, 2013), and slightly revised in 1988. Note that the 2013 Rural-Urban Continuum Codes are not directly comparable with the codes prior to 2000 because of the new methodology used in developing the 2000 metropolitan areas. See the Documentation for details and a map of the codes. |
| **Urban\_influence\_code\_2013** | The 2013 Urban Influence Codes form a classification scheme that distinguishes metropolitan counties by population size of their metro area, and nonmetropolitan counties by size of the largest city or town and proximity to metro and micropolitan areas. The standard Office of Management and Budget (OMB) metro and nonmetro categories have been subdivided into two metro and 10 nonmetro categories, resulting in a 12-part county classification. This scheme was originally developed in 1993. This scheme allows researchers to break county data into finer residential groups, beyond metro and nonmetro, particularly for the analysis of trends in nonmetro areas that are related to population density and metro influence. |
| **Civilian\_labor\_force\_2011** | 2013 civilian labor force estimate from US census |
| **Median\_Household\_Income\_2013** | 2013 Median household income estimate US census |
| **Median\_Household\_Income\_Percent\_of\_State\_Total\_2013** | 2013 Percent of State total for Median household income estimate US census |
| **Pct\_less\_than\_hs\_2009\_2013** | Percent of adults who did not graduate college 2009-2013 estimate from census |
| **Pct\_HS\_Grad\_only\_2009\_2013** | Percent of adults who only graduated high school 2009-2013 estimate from census |
| **Pct\_SomeCollege\_2009\_2013** | Percent of adults who attended some college 2009-2013 estimate from census |
| **Pct\_college\_grad\_2009\_2013** | Percent of adults graduated college or more 2009-2013 estimate from census |
| **Net\_Mig\_2010\_2014** | Calculated from Census net migration data from 2010 to 2014 |
| **MigRate\_2010\_2014** | Calculated from Census net migration data from 2010 to 2014 |
| **PctBlack** | Downloaded from Census |

**4.1.2 Economic Variables**

Data on the number and types of business establishments was downloaded from the US census and data for the amount of stimulus each county received was downloaded from a propublica research project at https://projects.propublica.org/recovery/.

|  |  |
| --- | --- |
| **Total\_Business\_Establishments\_2011** | 2011 census of busines (SUSB) estimate of total number of establishments by county |
| **Business\_Estab\_under\_20\_emp** | 2011 census of busines (SUSB) estimate of number of establishments with under 20employees by county |
| **Business\_Estab\_20to99\_emp** | 2011 census of busines (SUSB) estimate of number of establishments with 20 - 99 employees by county |
| **Business\_Estab\_100to499\_emp** | 2011 census of busines (SUSB) estimate of number of establishments with 100 - 499 employees by county |
| **Business\_Estab\_over500\_emp** | 2011 census of busines (SUSB) estimate of number of establishments with over 500 employees by county |
| **Stimulus** | - $$ in stimulus recieved by each county in US, pulled data from individual state pages on https://projects.propublica.org/recovery/ |
| **StimulusPerCapita** | -- calculated from data pulled from state pages on https://projects.propublica.org/recovery/ |
| **Stim\_PC\_Limit** | because of very large outliers in data I created a feature which limited the Stimulus Per Capita to $10,000 per person |

**4.1.3 Industry Typology Codes**

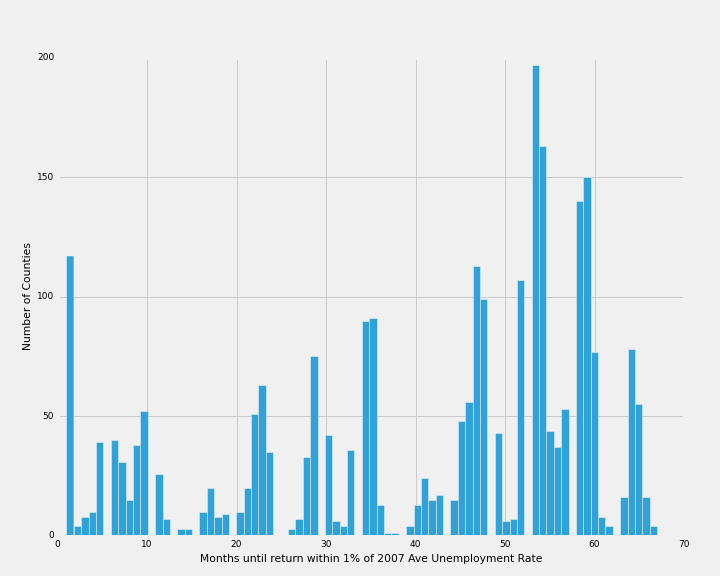
**Industrial Type Codes:** [**http://www.ers.usda.gov/data-products/county-typology-codes.aspx**](http://www.ers.usda.gov/data-products/county-typology-codes.aspx)

“The 2004 County Typology codes classify all U.S. counties according to six non-overlapping categories of economic dependence and seven overlapping categories of policy-relevant themes. The economic types include farming, mining, manufacturing, services, Federal/State government, and unspecialized counties. The policy types include housing stress, low education, low employment, persistent poverty, population loss, nonmetro recreation, and retirement destination. In addition, a code identifying counties with persistent child poverty is available.”

|  |  |
| --- | --- |
| **farm** | Farm-dependent county indicator. 0=no 1=yes |
| **mine** | Mining-dependent county indicator. 0=no 1=yes |
| **manf** | Manufacturing-dependent county indicator. 0=no 1=yes |
| **fsgov** | Federal/State government-dependent county indicator. 0=no 1=yes |
| **serv** | Services-dependent county indicator. 0=no 1=yes |
| **nonsp** | Nonspecialized-dependent county indicator. 0=no 1=yes |
| **house** | Housing stress county indicator. 0=no 1=yes |
| **loweduc** | Low-education county indicator. 0=no 1=yes |
| **lowemp** | Low-employment county indicator. 0=no 1=yes |
| **perpov** | Persistent poverty county indicator. 0=no 1=yes |
| **poploss** | Population loss county indicator. 0=no 1=yes |
| **rec** | Nonmetro recreation county indicator. 0=no 1=yes |
| **retire** | Retirement destination county indicator. 0=no 1=yes |
| **perchldpov** | Persistent child poverty county indicator ( 0=no 1=yes). This code identifies counties in which the poverty rate for related children under 18 years old was 20% or more in 1970, 1980, 1990, and 2000. This indicator was not part of the original ERS county typology codes and was added to this file after several requests. There are a total of 734 persistent child poverty counties. Of the 734 persistent child poverty counties, 603 are nonmetro counties, 131 are metro. The persistent child poverty codes were added March 17, 2009. |

**4.1.4 Seasonal Feature Engineering**

While exploring my data I noticed a seasonal trend in my data, with more counties returning to pre-recession rate during certain months. See graph:



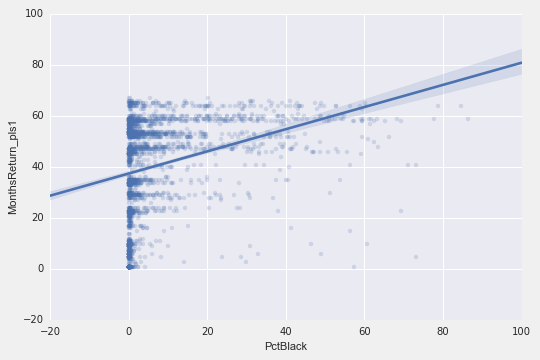
|  |  |
| --- | --- |
| **Summer678** | calculated a seasonal binary code for summer based on a county returning to within 1% of its prerecession unemployment rate in June, July, or August |
| **Spring345** | calculated a seasonal binary code for summer based on a county returning to within 1% of its prerecession unemployment rate in March, April, or May |
| **Fall91011** | calculated a seasonal binary code for summer based on a county returning to within 1% of its prerecession unemployment rate in September, October, or November |
| **HolidaySeason** | calculated a seasonal binary code for summer based on a county returning to within 1% of its prerecession unemployment rate in November or December |

**4.2 Feature Exploration**

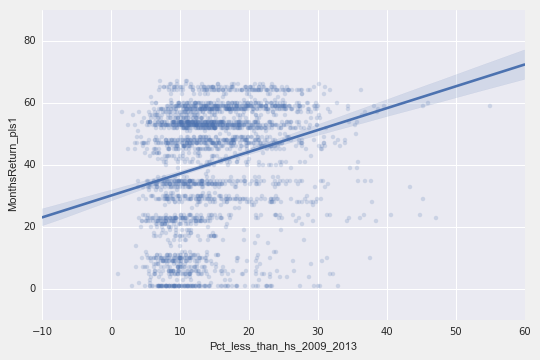
Using seaborn and matplotlib I created several scatterplots, map scatterplots, boxplots, histograms and a correlation matrix in order to look at the variables in question.

**4.2.1 Scatterplots**

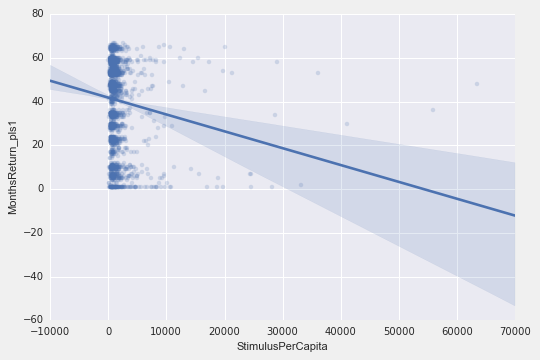
**Percent Black** -- Positive relationship between percent African American and the number of months until a county returned to its pre-recession unemployment rate.



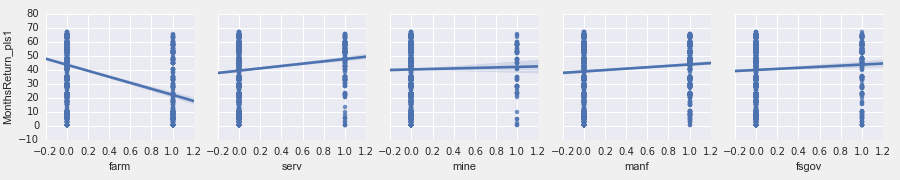
**Percent that did not graduate from high school** -- Positive relationship between percent that did not graduate from high school and the number of months until a county returned to its pre-recession unemployment rate.



**Stimulus Per Capita** -- Negative relationship between per capita stimulus and the number of months until a county returned to its pre-recession unemployment rate, but with significant outliers.

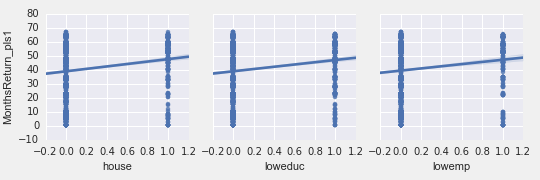


**Industrial Type Codes --** Positive relationship between service based economies and the number of months until a county returned to its pre-recession unemployment rate, negative relationship between farm-based economies and

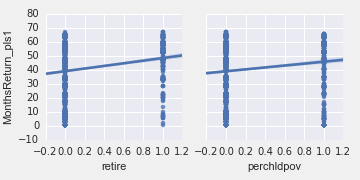


**Policy County Codes --**

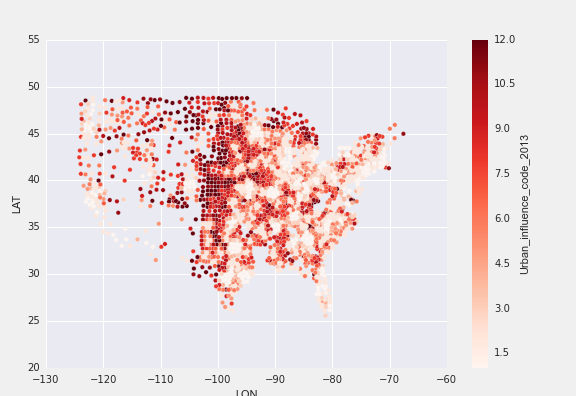
Many of the policy codes had useful differences.



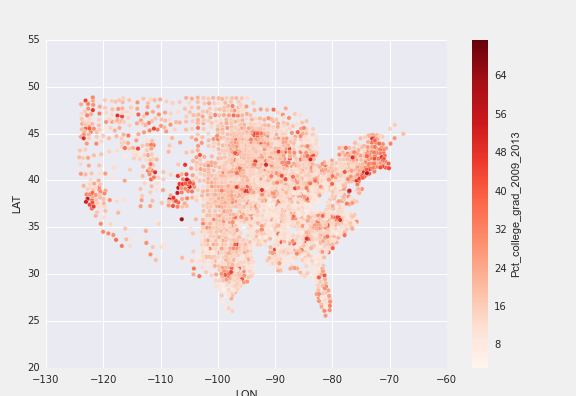




**Urban Influence Code** -- Urban influence code helps to capture the success that counties in the middle of the country had in returning to their pre-recession unemployment rates.

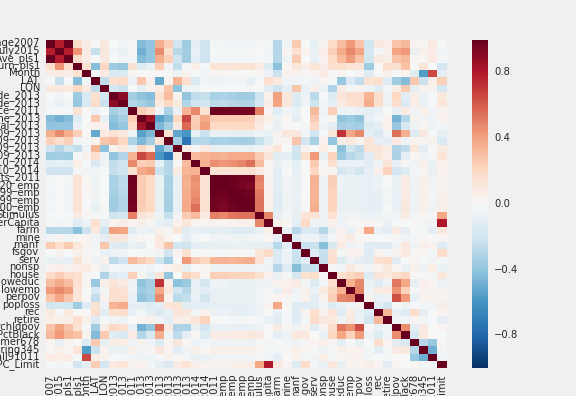


**Pct College Grad** -- Percent College Grad influence code helps to capture the success that counties in suburban counties.



**Example correlation matrix heatmap**

The correlation matrix helped me the most in picking features, as it helped identify the variables that we’re most correlated with our response variables.



**5. Modeling**

**5.1 Classification Modeling**

I chose logistic regression as my approach to understand whether a county will return to within 1% of its pre-recession unemployment rate or not. In order to do this I cleaned my dataset because several datasets were not available in all counties. I dropped any county that had null values in the industry types or the demographic features. I used notnull() to drop out any features and created a new data frame called cnty\_data\_notnull\_cl. I ended up with 2975 counties to test on.

I used train, test, split to create a training and testing subset and defined two functions that would return the AUC score and a confusion matrix for each iteration of the model.

The best model I found (so far) had an AUC of .94 but has problems in that it had an extremely skewed confusion matrix. I do better at true positives than true negatives.

Confusion Matrix:

array([[ 43, 44],

[ 11, 646]]

-- Working to try to fix this?

**5.2 Regression Modeling**

I chose linear regression as my approach to understand how many months it will take for a county will return to within 1% of its pre-recession unemployment rate. In order to do this I cleaned my dataset because several datasets were not available in all counties. I dropped any county that had null values in the returned to within 1%, industry types, or demographic features. I used notnull() to drop out any features and created a new data frame called cnty\_data\_notnull\_cl. I ended up with 2631 counties to research on.

I used train, test, split to create a training and testing subset and defined a function that would return the RMSE score for each iteration of the model.

I calculated the null RMSE by using the average months it took until a county recovered and it was 18.631.

The best model iteration RMSE was 13.7 which is better than the null RMSE but I hope to find a better model.

**6. Findings and Conclusions**

Still in work.