Migration and Credit Insecurity

By: David Woolf, 500933541

***Initial preparation:***

The original CII data was close to what was eventually used for data regression. The Migration data required a great deal more cleaning. Each State had its own Excel sheet that had to be consolidated in order to upload. In addition, the first two year 2005-2009, and 2006-2010, were not formatted like the later years. Through careful Excel craft, those were made to be like the other years. However, the margins could not be reproduced, as such they were left to equal their corresponding estimates. Part of the preparation was creating the “Transit Route,” a single, unique number, that denotes the State and County being moved to, followed by the State and County being moved from, using the FIPS system to encode this information. After cleaning formatting each set, they were all uploaded to R.

***Data Craft:***

The datasets was then joined together on “Transit Route” as unique identifier, to form 1 large data set. This ended up being a very large data set with 758,838 rows. The joined data set, all had a State and County being moved to, and a State and County being moved from, but not in a unified column. Unable to resolve this issue in R, I reimported selection of data to Excel, and used an Excel formula to find the longest string in a county or state (all states and counties have a longer length than NA), and put that into a new column. Once done, the state and county information were now consolidated, and able to be reuploaded to R for further processing.

The next stage was combining the CII with the migration data set. Reimporting to Excel, and using Index and Match in Excel, every county gets 3 corresponding CII values, 2007, 2012, and 2018. Once reupload to R, the pruning of not valuable data is ready to occur. First, any County that did not appear in the CII was filtered out, this includes all of Puerto Rico. Next, immigration from outside of the US is removed. Having been completed, every row in the databases had a 3 corresponding CII values.

The initial plan was to run a linear model using this data set, however, R advised of the need for approximately 100 GB of memory to accomplish this. Attempts to partition the data, into smaller sizes (attempted with 50 subsamples), did not enable the memory limitations to be overcome. Unless a work around is achieved or greater technology is accessed, this route is inaccessible at this time.

The alternative was filtering to a specific state, Alabama, to reduce the data size enough for regression. The originally, both immigration and emigration were going to be used for modeling, however, that has proven to be too resource intensive to be complete at this time. Gross migration was selected over net migration as it is strictly positive, which avoid certain coding issues. The first model is the training model, trying to predict the 2012 index, using the 2007 index, and Gross migration of years between 2005 and 2012. The other model created was the test model, trying to predict the 2018 index using the 2012 index, and Gross migration from 2009 to 2017. Both models were able to but run, and it will just require more time to run for the other 49 States.