Project Data Report

Lorenz Menendez

Introduction

Isaac and I separated the data wrangling into two equal parts. Isaac worked on cleaning and wrangling election result data, which he will describe in his data report. I worked on wrangling county boundary data with demographic and social data. Here are the input datasets I worked with.

Input Datasets

Name	*Description*	*Variables*	*Temporal Resolution*	*Spatial Resolution*	*File Format*
County Facts	Demographic and social data for each county in the U.S.	FIPS code, Name, State, Population, + other race/gender/education/… variables.	2010-2014	County	.csv
County Boundaries	Multipolygon GIS data representing county boundaries	State & County FIPS code, GEOID, Name, Land Area, Water Area	2016	County	.shp

I developed an R script to automatically to download, process and import clean data into R. You can view it here. The R script takes the input datasets and joins them together based on FIPS codes to create a master dataset called 'electnData16'.

I tried my best to have R import from URLs so that it's very reproducible. Unfortunately, using a simple URL is not working for the County Facts dataset. I think I'll have to eventually setup a Kaggle API to automatically import that data into R.

Output Datasets

Running a spatial autocorrelation algorithm comparing each individual demographic variable with election results will create a new dataset with correlation coefficients for each variable. Depending on the results, this could be out final dataset. A further step we could take would be to find correlations between the first-order correlations to create second-order correlations.

Sources

County Facts

Hamner, Ben. "2016 US Election." Kaggle, 1 July 2016, www.kaggle.com/benhamner/2016-us-election.

County Boundaries

US Census Bureau. "Cartographic Boundary Files - Shapefile." Cartographic Boundary Files - Shapefile, 2 May 2019, www.census.gov/geographies/mapping-files/time-series/geo/carto-boundary-file.html. Downloaded: cb_2018_us_county_500k.zip

Importing data from Kaggle and the U.S. Census

County Facts — *countyQuickfacts16*		
PK	fips	integer
	area_name	string
	state_abbreviation	string
	PST045214	integer
	PST040210	integer
	POP010210	integer
	AGE135214	integer
	+ 50 Other Demographic Variables	integer

County Boundaries — *countyBoundaries16*		
	STATEFP	integer
	COUNTYFP	integer
PK	COUNTYNS	integer
PK	GEOID	integer
	NAME	string
	LSAD	integer
	ALAND	integer
	AWATER	integer
	Geometry	geometry

New Calculated Variables

County Correlation — *countyCorrelations16*		
PK	fips	integer
	area_name	string
	state_abbreviation	string
	PST045214_Corr	numeric
	PST040210_Corr	numeric
	POP010210_Corr	numeric
	AGE135214_Corr	numeric
	+ 50 Other Demographic Variable Correlations	numeric