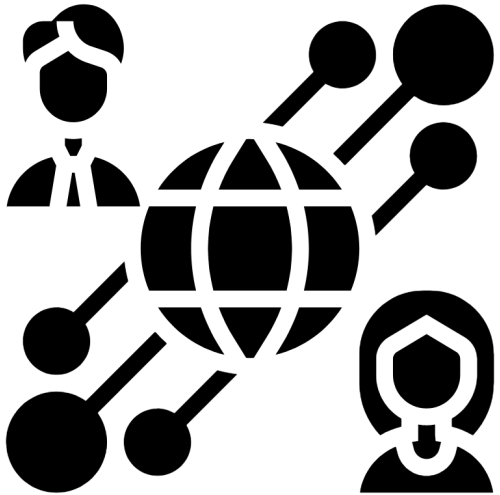
Household Income by U.S. County

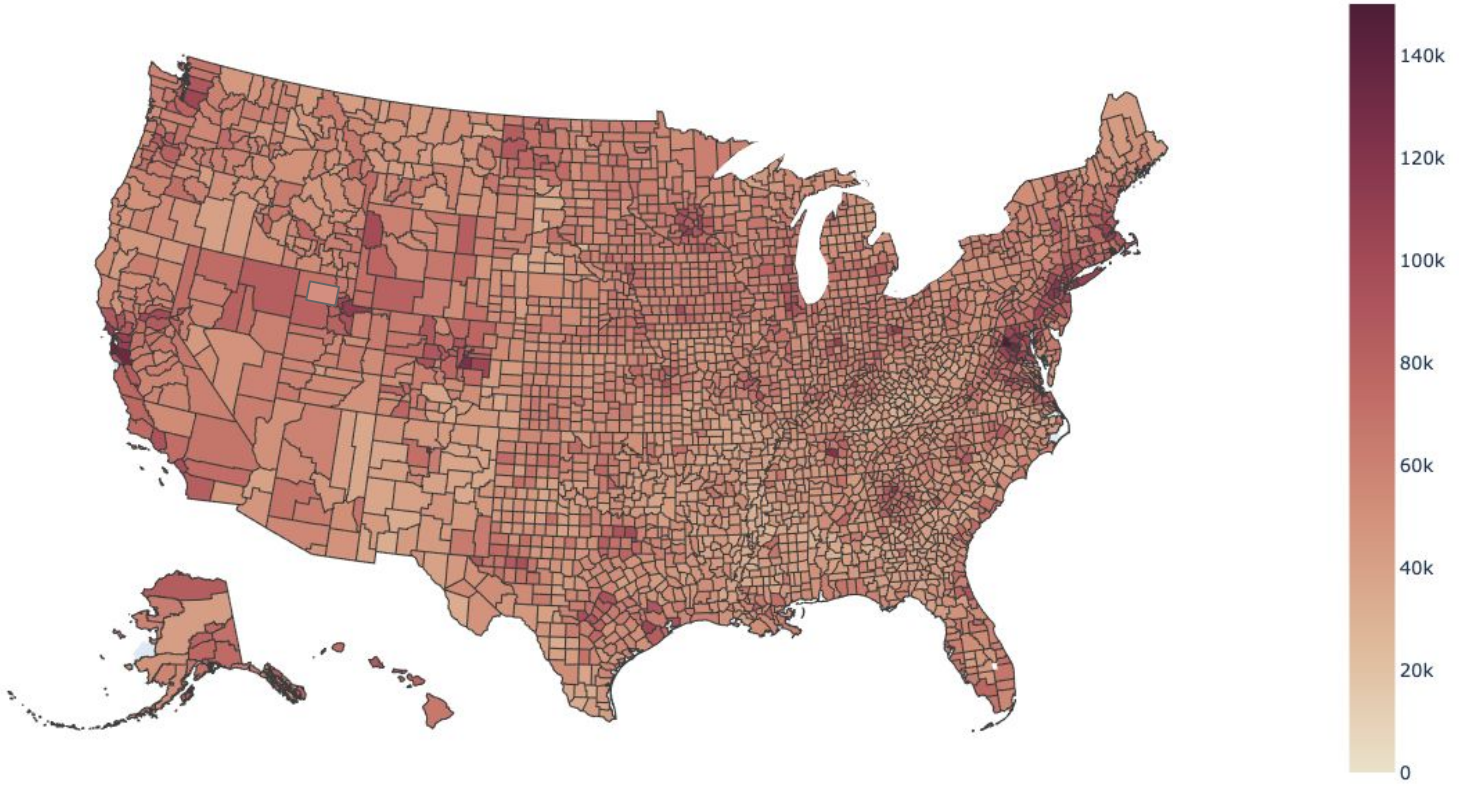
Predicting Median Household Income with Linear Regression

Amy Butler

Introduction



2019 Median Household Income



Methodology

Observations: 3,193 counties and county-equivalents

Data: Scraped from Wikipedia pages and files from USDA

Model: Linear Regression with Ridge Regularization

Data Gathered

Median Household Income

Population

Civilian Labor Force

Employed

Unemployed

Unemployment Rate

Number of major highways

Total number of cities and towns

Data Gathered

Median Household Income

Population

Civilian Labor Force

Employed

Unemployed

Unemployment Rate

Number of major highways

Total number of cities and towns

Economic Type

- **Non-specialized**
- **Recreation**
- **Manufacturing**
- **Farming**
- **Mining**
- **Federal/State Government**

Data Gathered

Median Household Income

Population

Civilian Labor Force

Employed

Unemployed

Unemployment Rate

Number of major highways

Total number of cities and towns

Economic Type

Rural Urban Continuum Classification

1. **Metro - Counties in metro areas of 1 million population or more**
2. **Metro - Counties in metro areas of 250,000 to 1 million population**

...

8. **Nonmetro - Completely rural or less than 2,500 urban population, adjacent to a metro area**
9. **Nonmetro - Completely rural or less than 2,500 urban population, not adjacent to a metro area**

Data Gathered

Median Household Income

Population

Civilian Labor Force

Employed

Unemployed

Unemployment Rate

Number of major highways

Total number of cities and towns

Economic Type

Rural Urban Continuum Classification

Retirement Destination

Low Education County

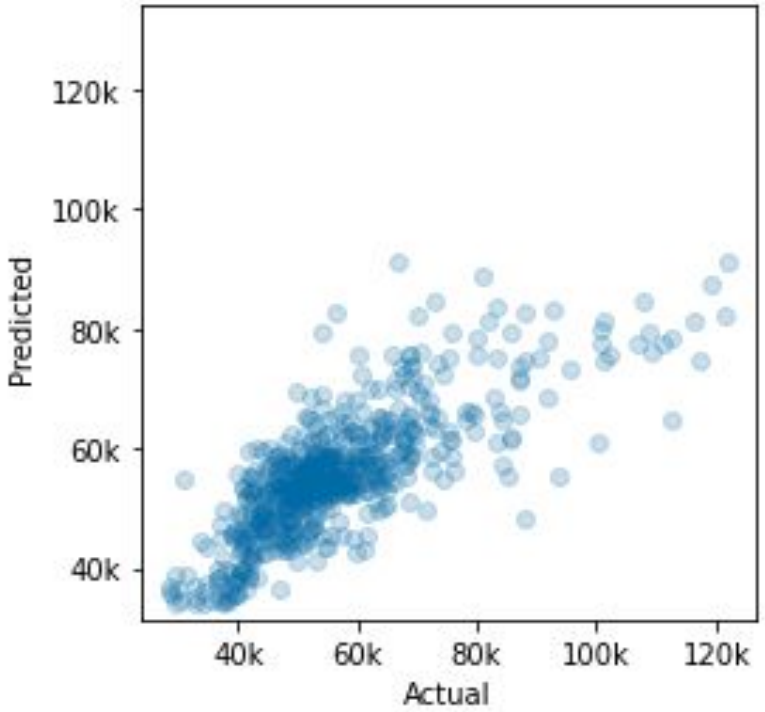
Low Employment County

Persistent Poverty

Results

Model	R^2	RMSE
Linear Regression with Ridge Regularization	**64%**	**$9,172**

Results



Important Features

- Unemployment Rate
- Civilian Labor Force
- Persistent Poverty
- Low Employment County

Excluded Features

- County Population
- Employed
- Unemployed
- Number of major highways

Conclusions

1. There is overlap in the median household income ranges of metro and non-metro counties.

Conclusions

1. There is overlap in the median household income ranges of metro and non-metro counties.

2. **The features included explained 64% of the variance in median household income across all U.S. counties.**

Conclusions

1. There is overlap in the median household income ranges of metro and non-metro counties.

2. The features included explained 64% of the variance in median household income across all U.S. counties.

3. **The RMSE would need to be decreased for the model to be more meaningful and applicable.**

Future Work

Include county resident demographic information

Explore options to capture under-employed population

Try a tree-based algorithm

Thank you!

Appendix

Predicting Median Household Income with Linear Regression

Data Sources

Median household income: U.S. Department of Commerce, Bureau of the Census, Small Area Income and Poverty Estimates (SAIPE) Program

Wikipedia: County(United States), pages for each State listing counties in the state, pages specific to each county

Tables prepared by USDA, Economic Research Service available at: https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/

Unemployment: U.S. Department of Labor, Bureau of Labor Statistics, Local Area Unemployment Statistics (LAUS)

Rural Area Continuum Codes

Metropolitan Counties

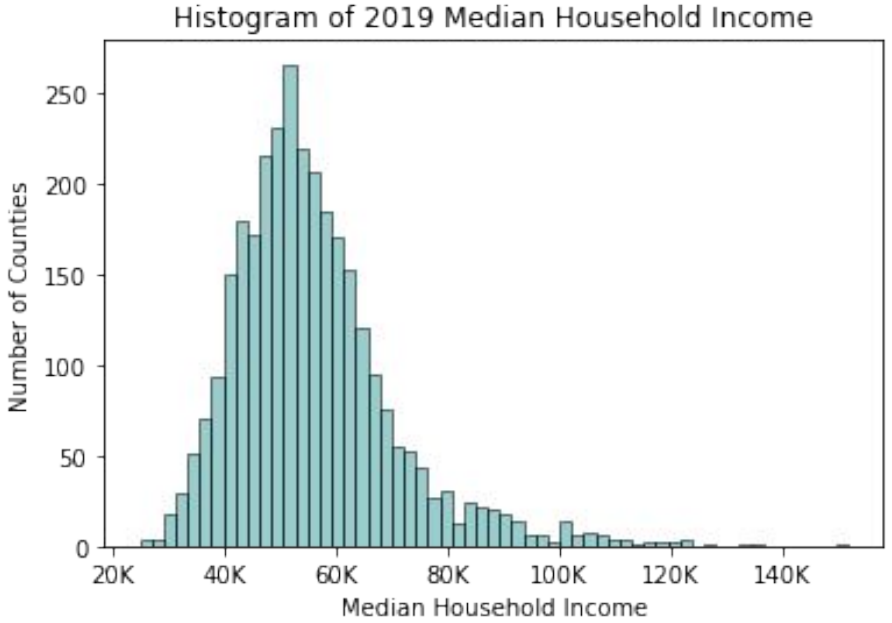
Code	Description
1	Counties in metro areas of 1 million population or more
2	Counties in metro areas of 250,000 to 1 million population
3	Counties in metro areas of fewer than 250,000 population

Nonmetropolitan Counties

Code	Description
4	Urban population of 20,000 or more, adjacent to a metro area
5	Urban population of 20,000 or more, not adjacent to a metro area
6	Urban population of 2,500 to 19,999, adjacent to a metro area
7	Urban population of 2,500 to 19,999, not adjacent to a metro area
8	Completely rural or less than 2,500 urban population, adjacent to a metro area
9	Completely rural or less than 2,500 urban population, not adjacent to a metro area

2019 Median Household Income

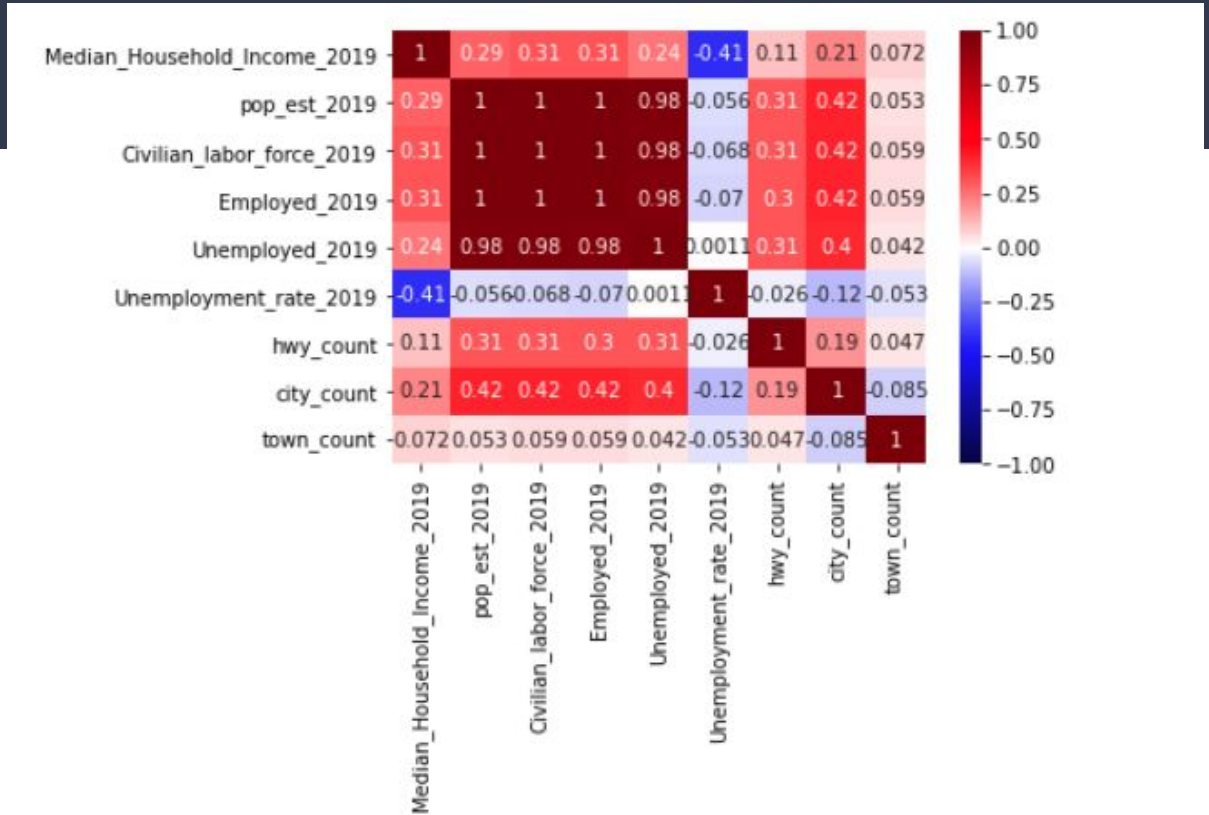
- Ranged from 25K to 150K
- 25th percentile: 46K
- 75th percentile: 62K



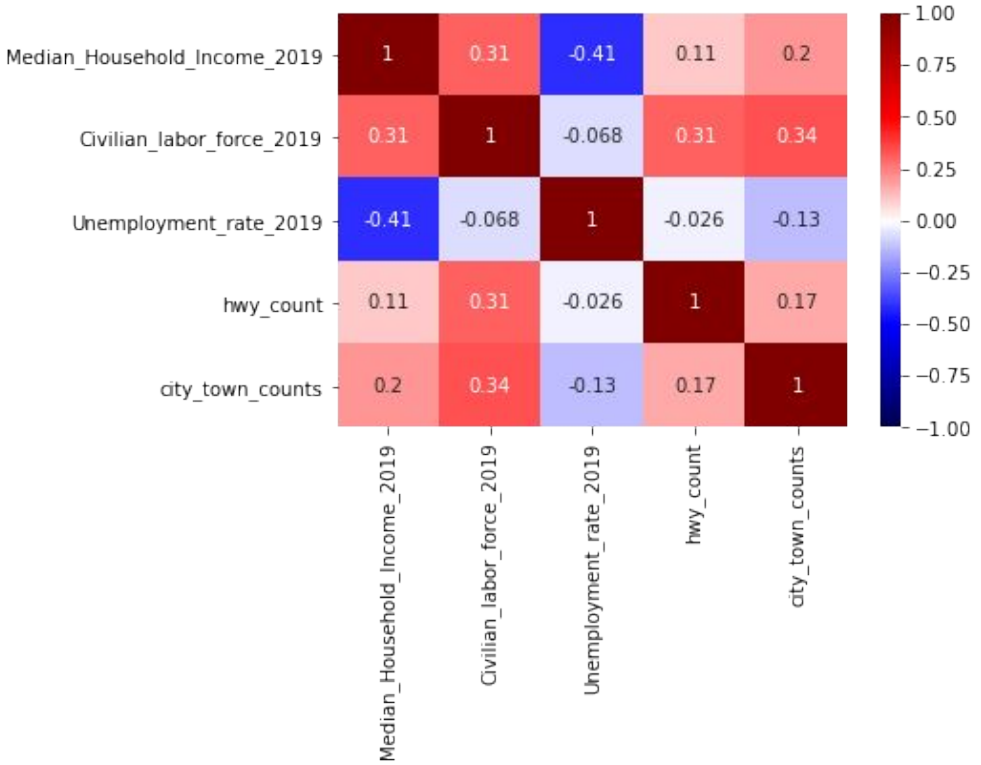
Correlations

Based on correlations, excluded from model fit:

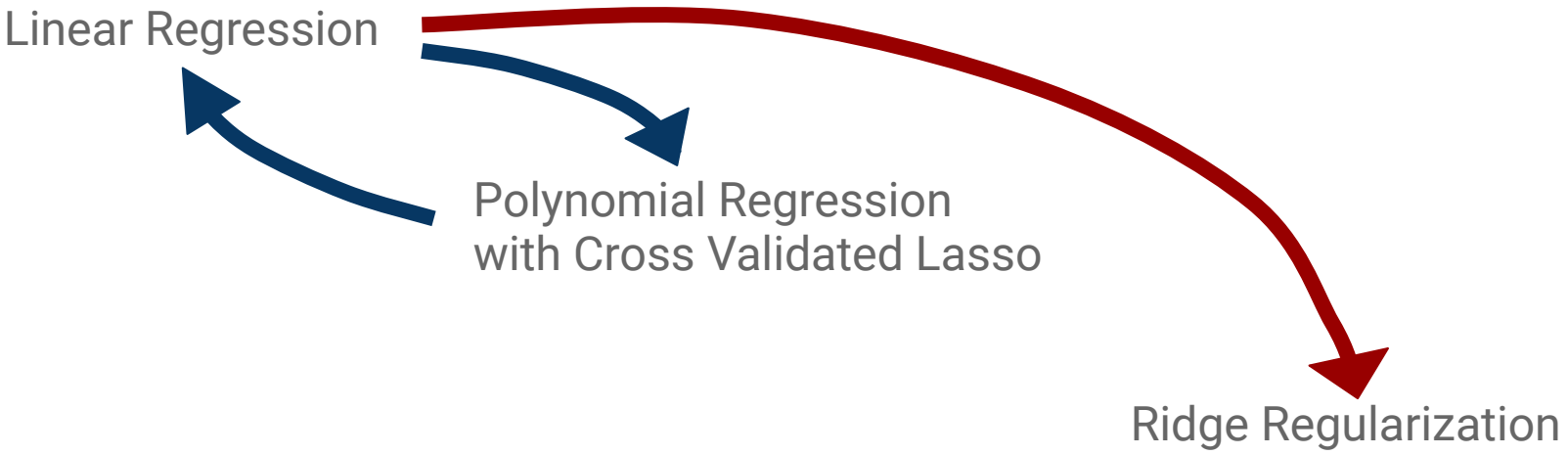
- County population
- Employed
- Unemployed



Correlation Matrix of Numeric Features Kept



Model Development



Final Model

Log transformed
Median Household Income =

```
('Farming_2015_Update', -0.010124187067450996),
('Government_2015_Update', -0.005046001152041013),
('Low_Education_2015_Update', -0.01833617620661021),
('Low_Employment_Cnty_2008_2012_25_64', -0.057395362886867585),
('Manufacturing_2015_Update', -0.008220746448404398),
('Mining_2015-Update', 0.008922405355270148),
('Nonspecialized_2015_Update', -0.009621967186791064),
('Persistent_Poverty_2013', -0.051885972961600024),
('Recreation_2015_Update', 0.002528843454841667),
('Retirement_Dest_2015_Update', 0.013509117883849397),
('Rural_urban_continuum_code_2013_1.0', -0.00596523075671801),
('Rural_urban_continuum_code_2013_2.0', 0.014121921093381166),
('Rural_urban_continuum_code_2013_3.0', 0.00880507276663978),
('Rural_urban_continuum_code_2013_4.0', -0.0006053486903045552),
('Rural_urban_continuum_code_2013_5.0', -0.004987495844668066),
('Rural_urban_continuum_code_2013_6.0', -0.003987291613158361),
('Rural_urban_continuum_code_2013_7.0', -0.0044355283153782454),
('Rural_urban_continuum_code_2013_8.0', -0.001128939385638572),
('Rural_urban_continuum_code_2013_9.0', -0.002573472857477749),
('Unemployment_rate_2019', -0.04604781842964534),
('city_town_counts', -0.0013367412393887612),
('log_Civilian_LF', 0.061991049166649825),
('log_civ_LF_sqrd', 0.01599392285016646),
('rec*civ_LF', 0.01016503216237205),
('rul*civ_LF', 0.0711927008243472),
('unemp*civ_LF', -0.08215707726023486),
('unemp_sqrd', 0.0834701810548302)]
```

Model Residuals