



Galvanize Denver g64

1º Capstone – Feb/16/2018
Nei Costa

Dataset: Census

Source: Fact Finder interface

<https://factfinder.census.gov>

Goal:

**Identify the factors that influence the percentage of senior in counties
(Seniors = people 60 years and over).**



This information could be used to improve health programs and other services specific to these people.

Imported files:

ACS (American Community Survey 5-Year Estimates)

Population Statistics: ACS_16_5YR_CP05_with_ann.csv

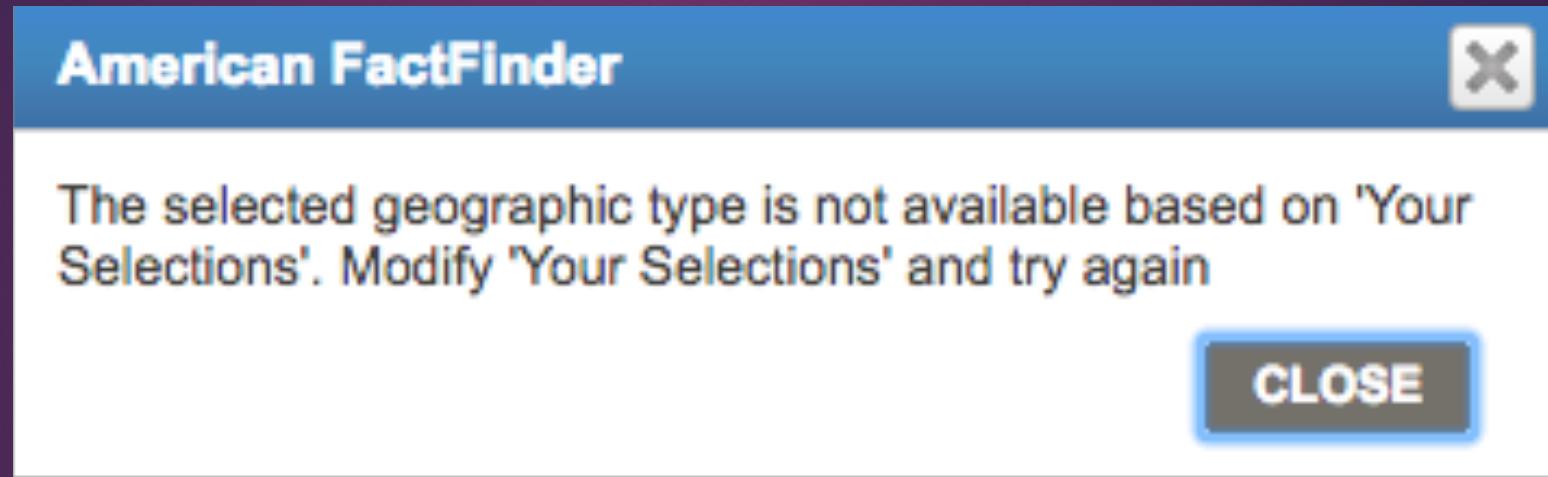
Per capita income : ACS_16_5YR_DP03_with_ann.csv

ECN (Economic Census of the United States)

Industries by County: ECN_2012_US_31A1_with_ann.csv

OBS: Unfortunately I did not find other information related to the weather, health condition, cost of living per County.

Information not available on this site:



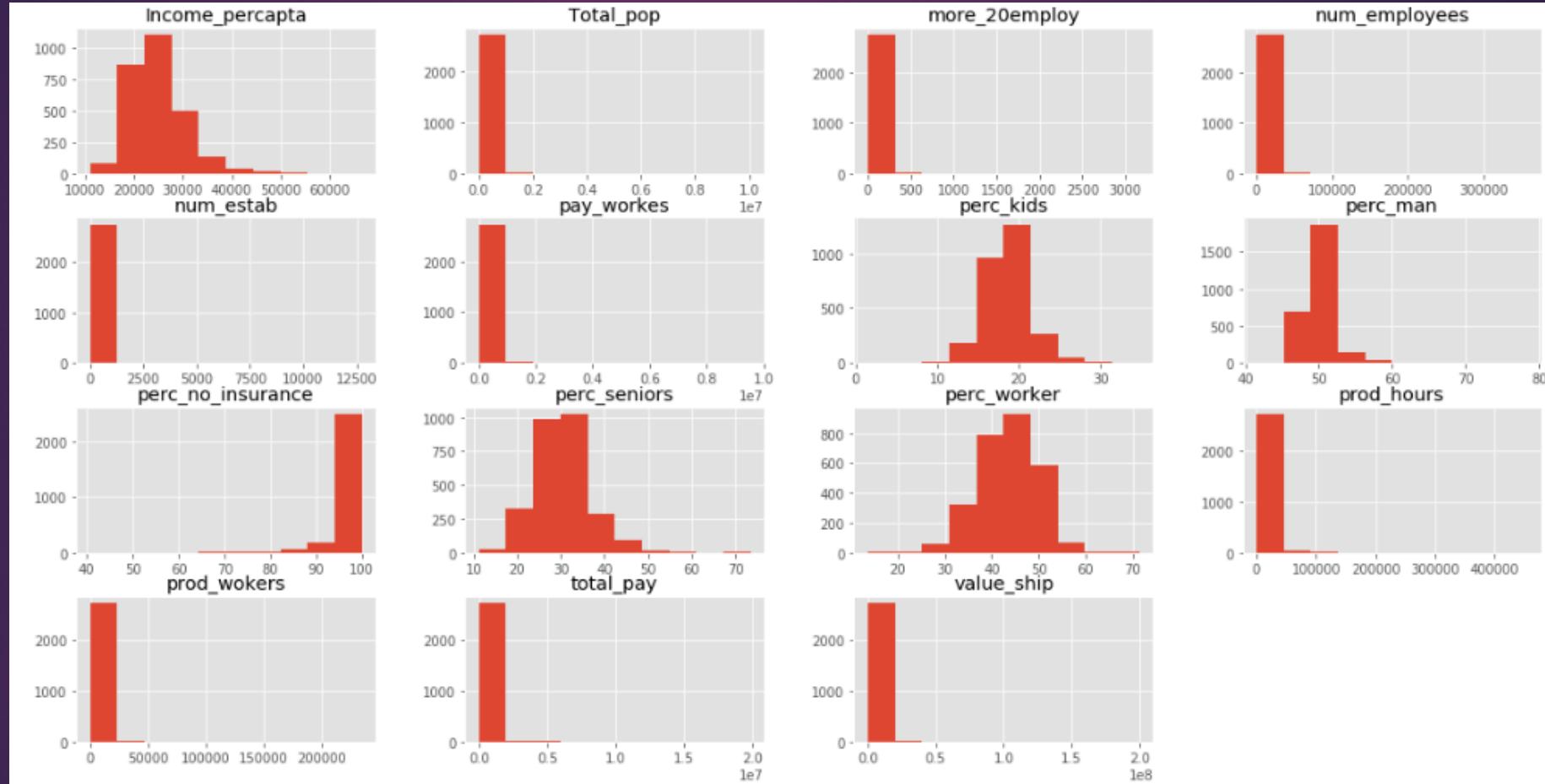
Some relevant information was not available by state or county.

Cleansing and Organizing Data:

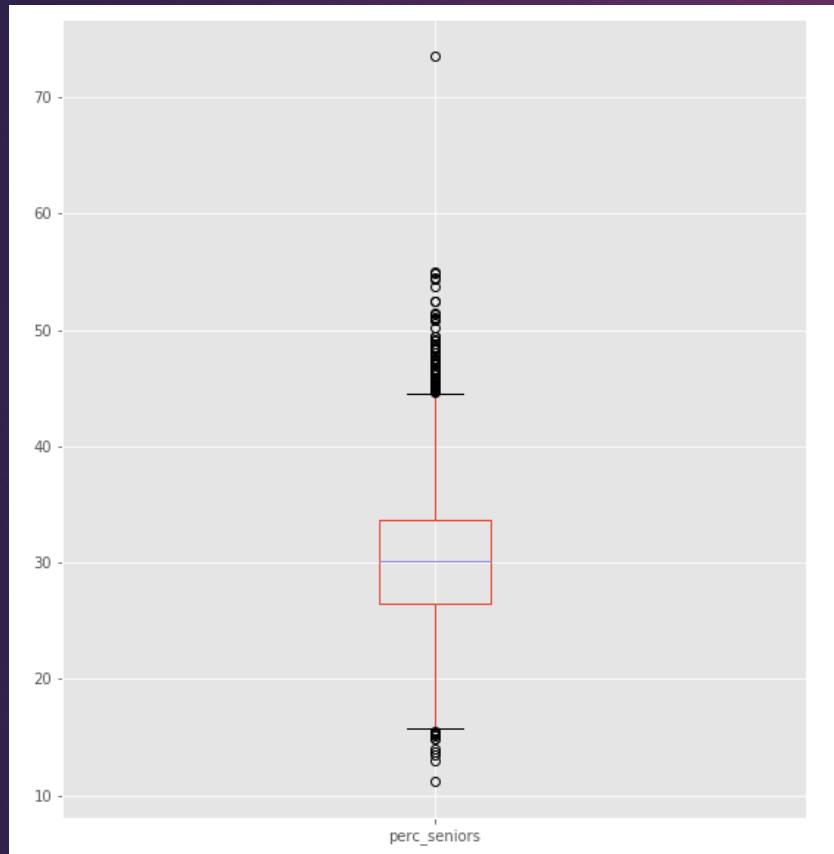
Imported files bring lots of information, requiring to delete unnecessary columns, and renaming columns to names most closely related to your content.

This work was done in the data select.py file in the SRC folder that provides the data already prepared.

Exploring the data:



Removing Outliers



Mean % Seniors in USA Counties: 30.33%
Std (6.13%)

More than 4 standard deviations

% Seniors < 5.81%:

% Seniors > 54.85%:

Charlotte (Florida): 55.00%

Sumter (Florida): 73.50%

Describe % Seniors

	perc_seniors
count	2760.00
mean	30.31
std	6.06
min	11.20
25%	26.50
50%	30.05
75%	33.70
max	54.80

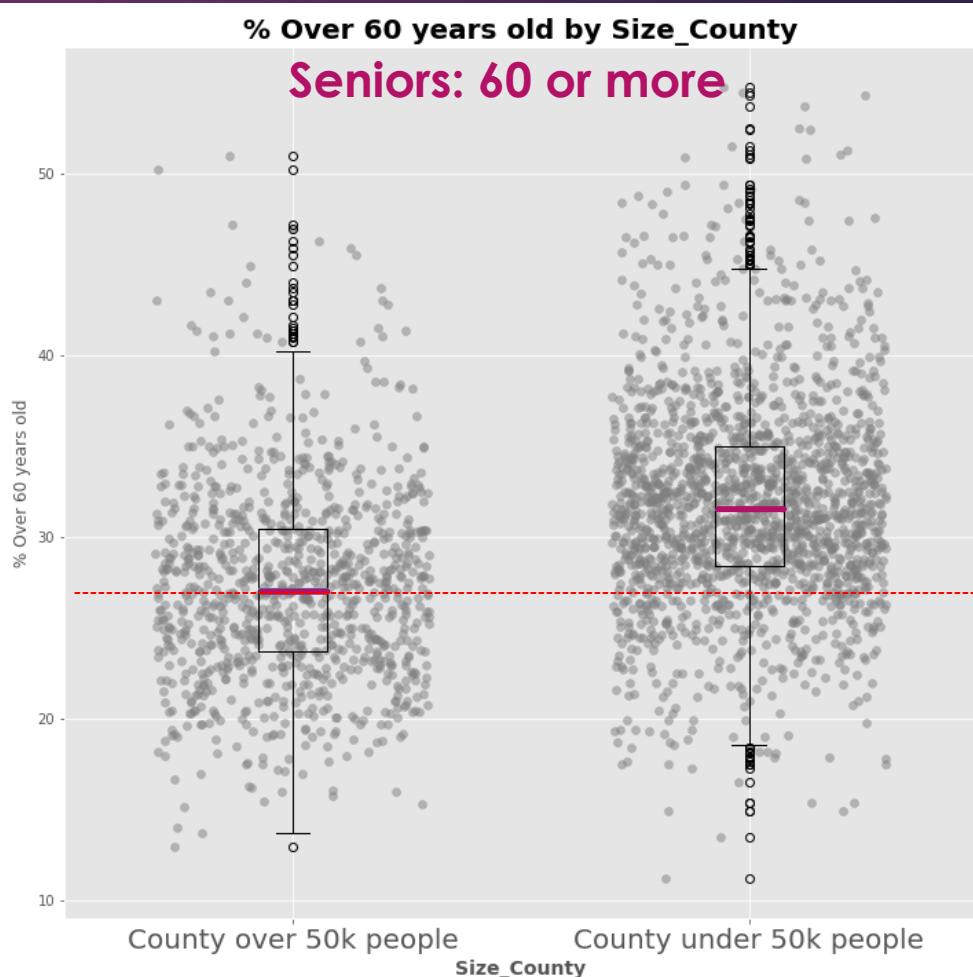
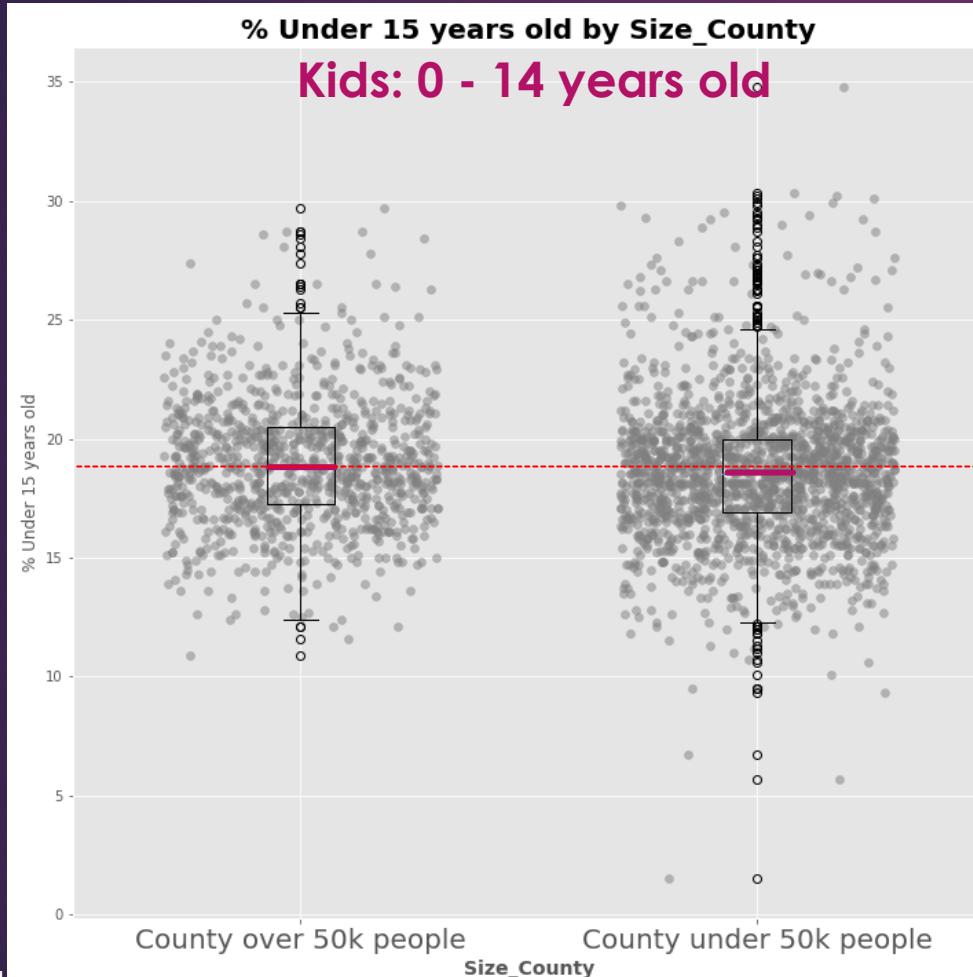
Comparing Counties large and small



Small County:
 < 50.000 people

Large County:
 > 50.000 people

Comparing Counties large and small



Comparing Counties large and small

Size_County	Total_pop	perc_man	perc_kids	perc_seniors
County over 50k people	284,384.9213	49.3496	18.9646	27.4079
County under 50k people	21,705.5730	50.2147	18.5466	31.8959

Using the t-test to confirm if this difference is statistically significant

H_0 : Mean % Seniors of Small County == Mean % Seniors of Big County

H_1 : Mean % Seniors of Small County != Mean % Seniors of Big County

t-stat: -20.2343 - P-val: 0.000

Reject H_0 :

Conclusion:

Mean % Seniors of Small County != Mean % Seniors of Big County

Why?

What factors could be causing the lowest percentage of seniors in the larger Counties?

I could find:

- Family income;
- Conditions of employment;

I couldn't find:

- Cost of living;
- Climate;
- Health conditions.

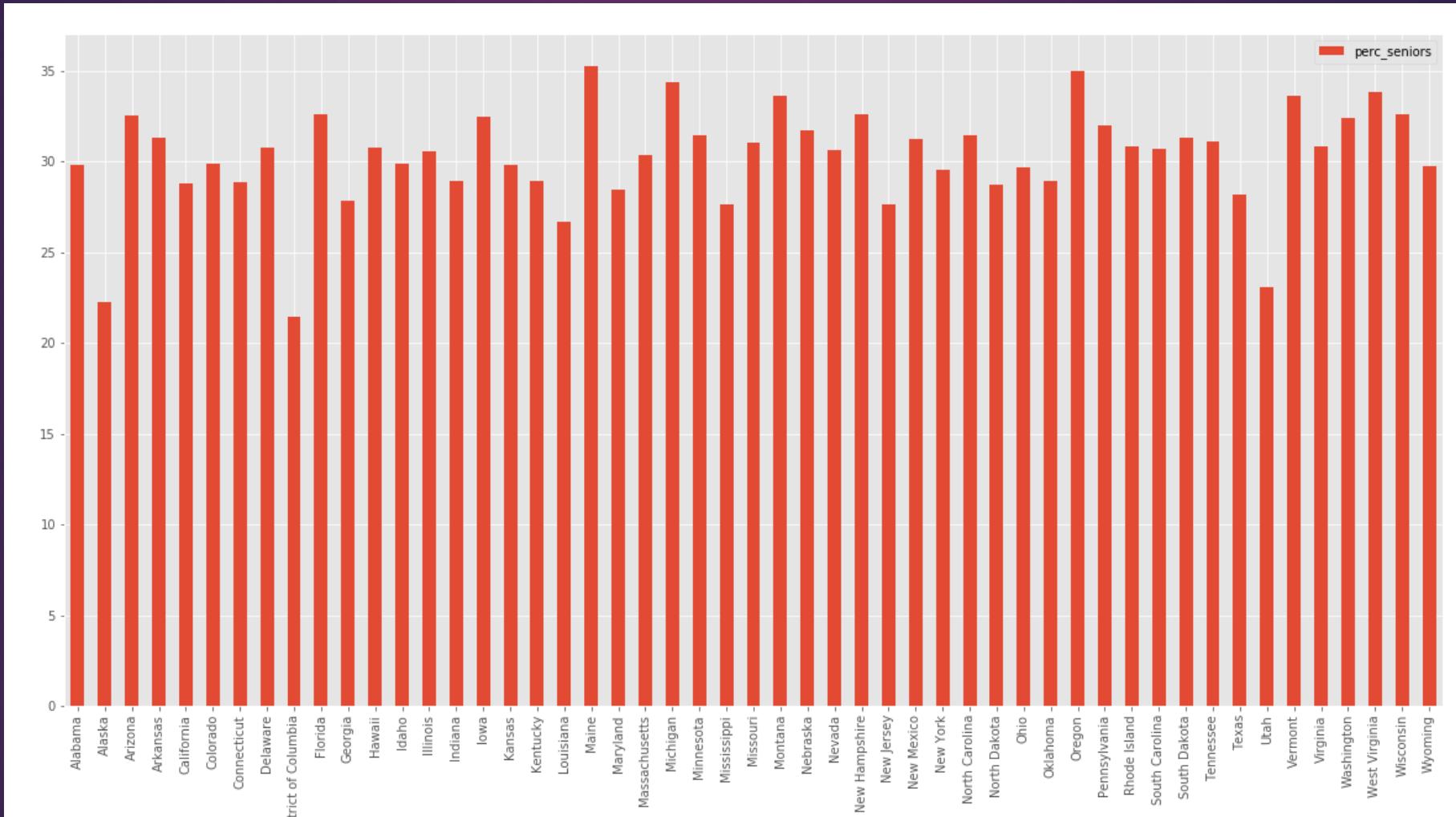


EDA

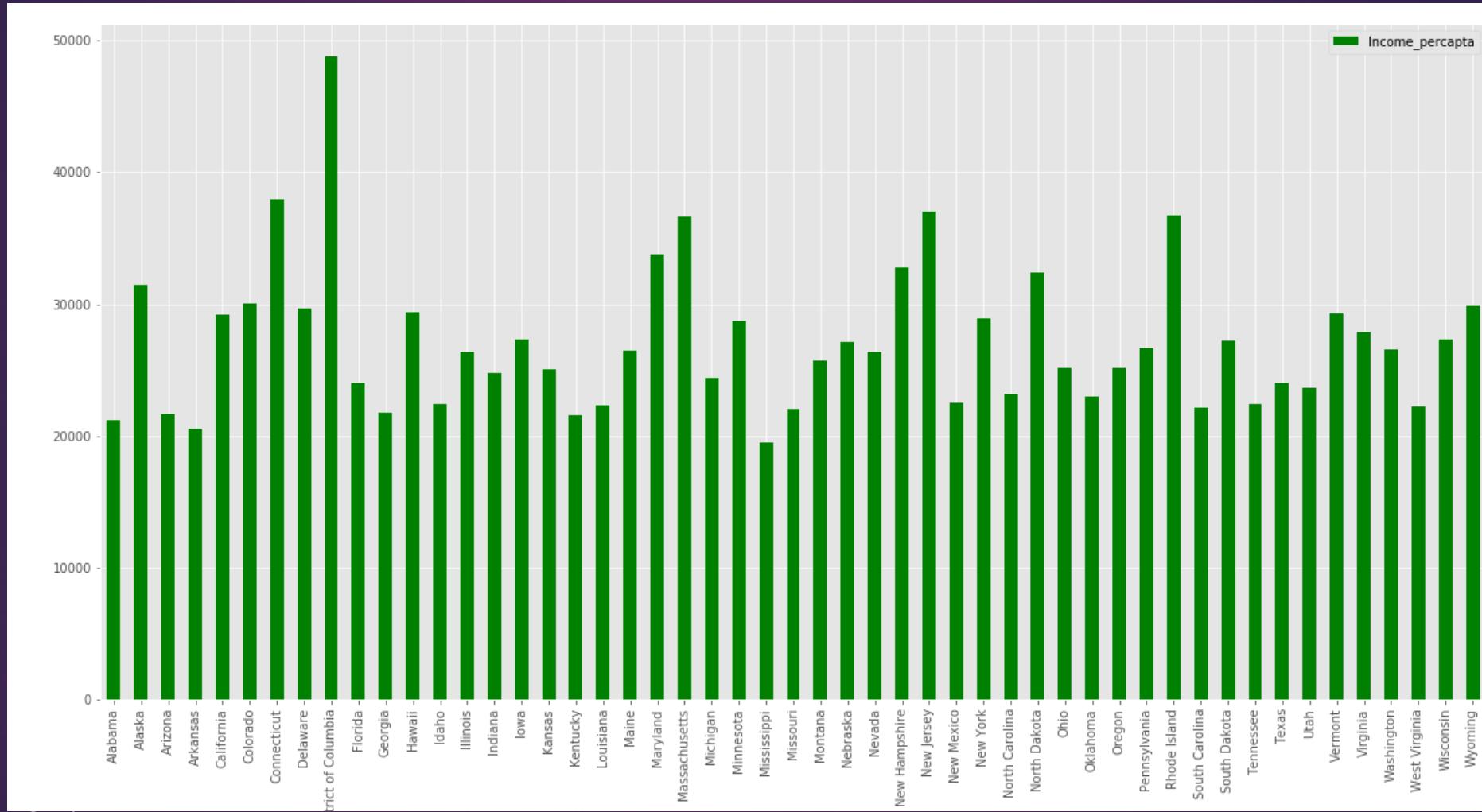
Well, let's work with what we have then.

Some information grouped
by state (to facilitate
visualization)

% Seniors

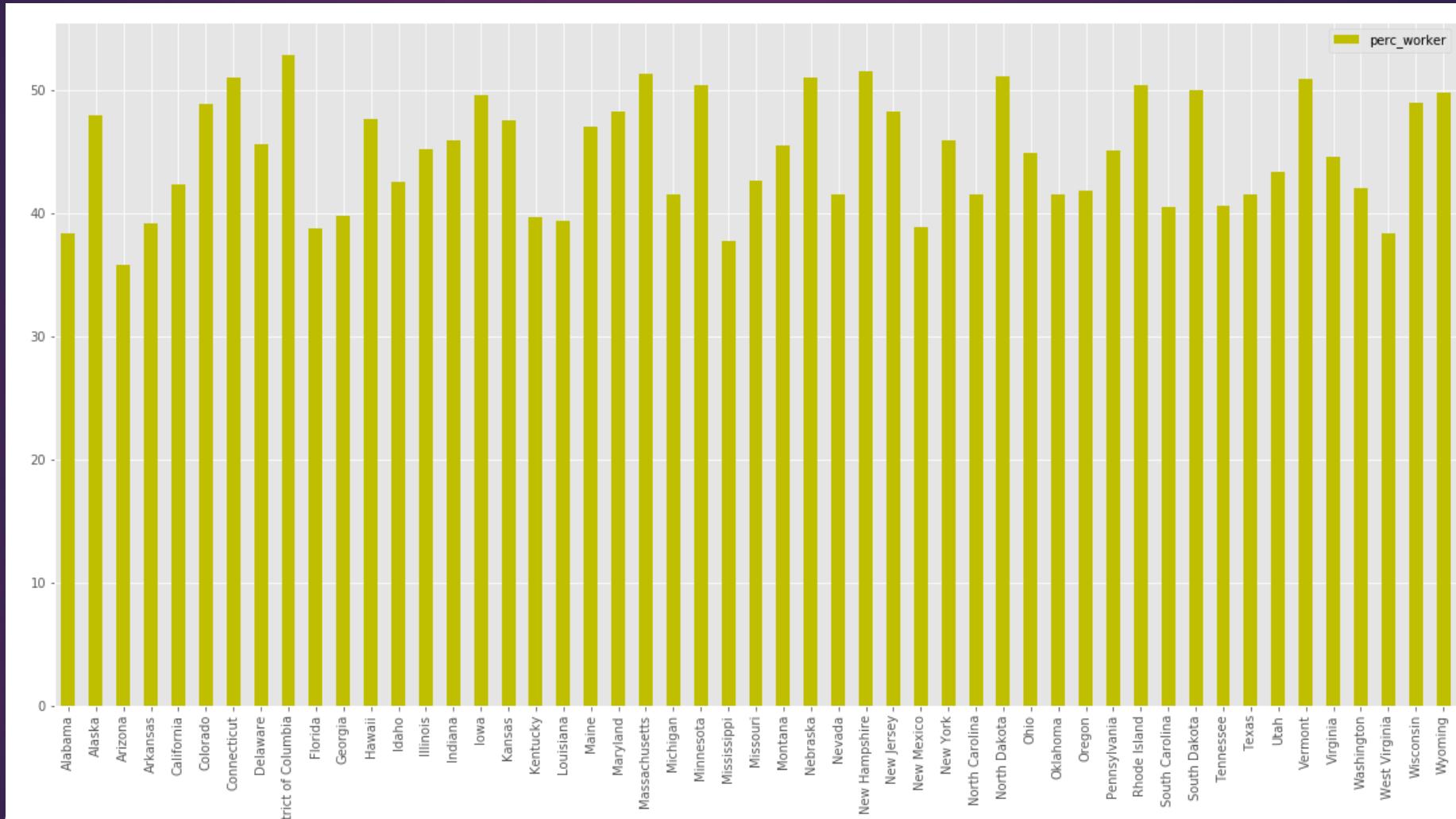


Per Capita Income

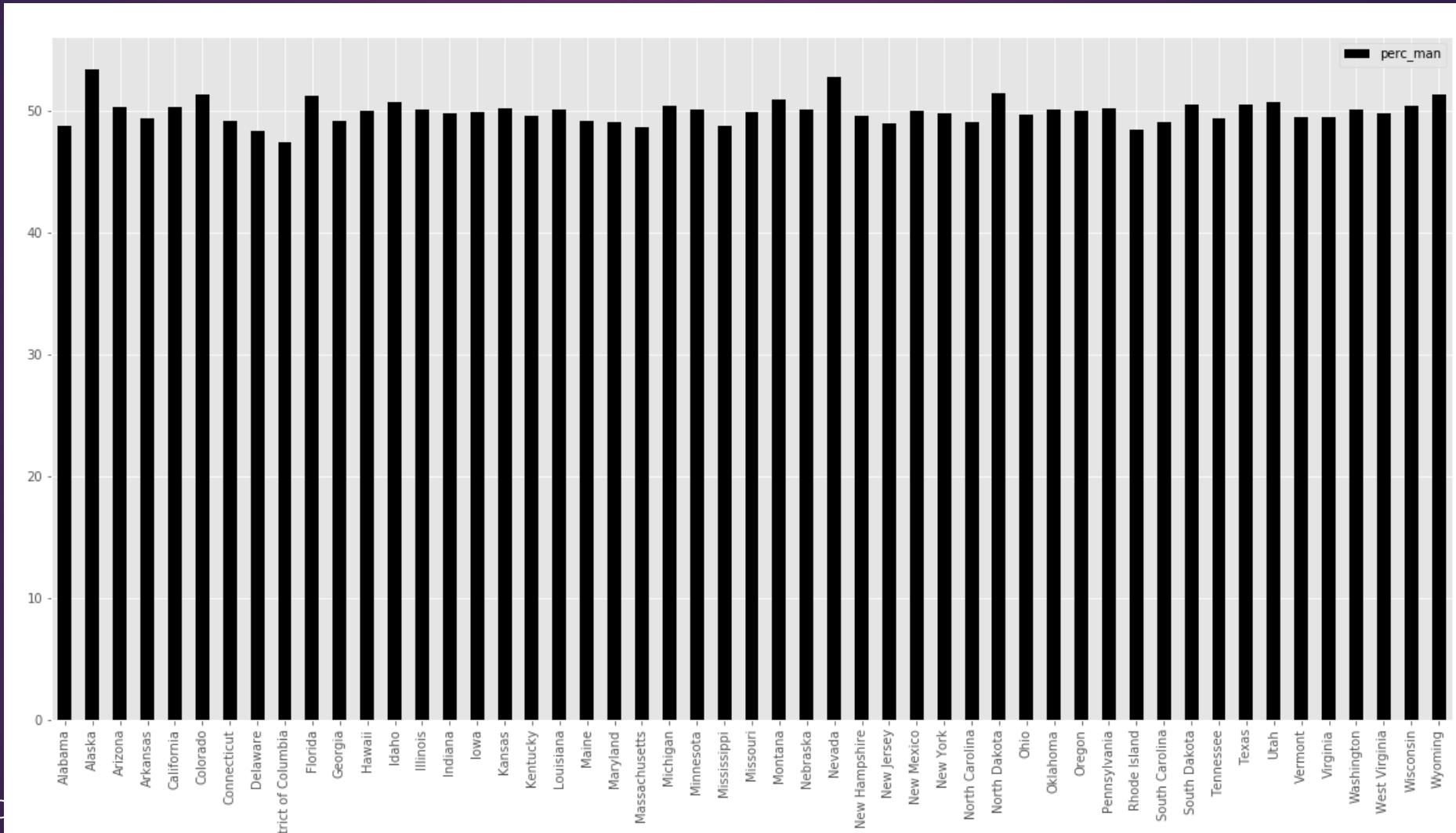


galvanize

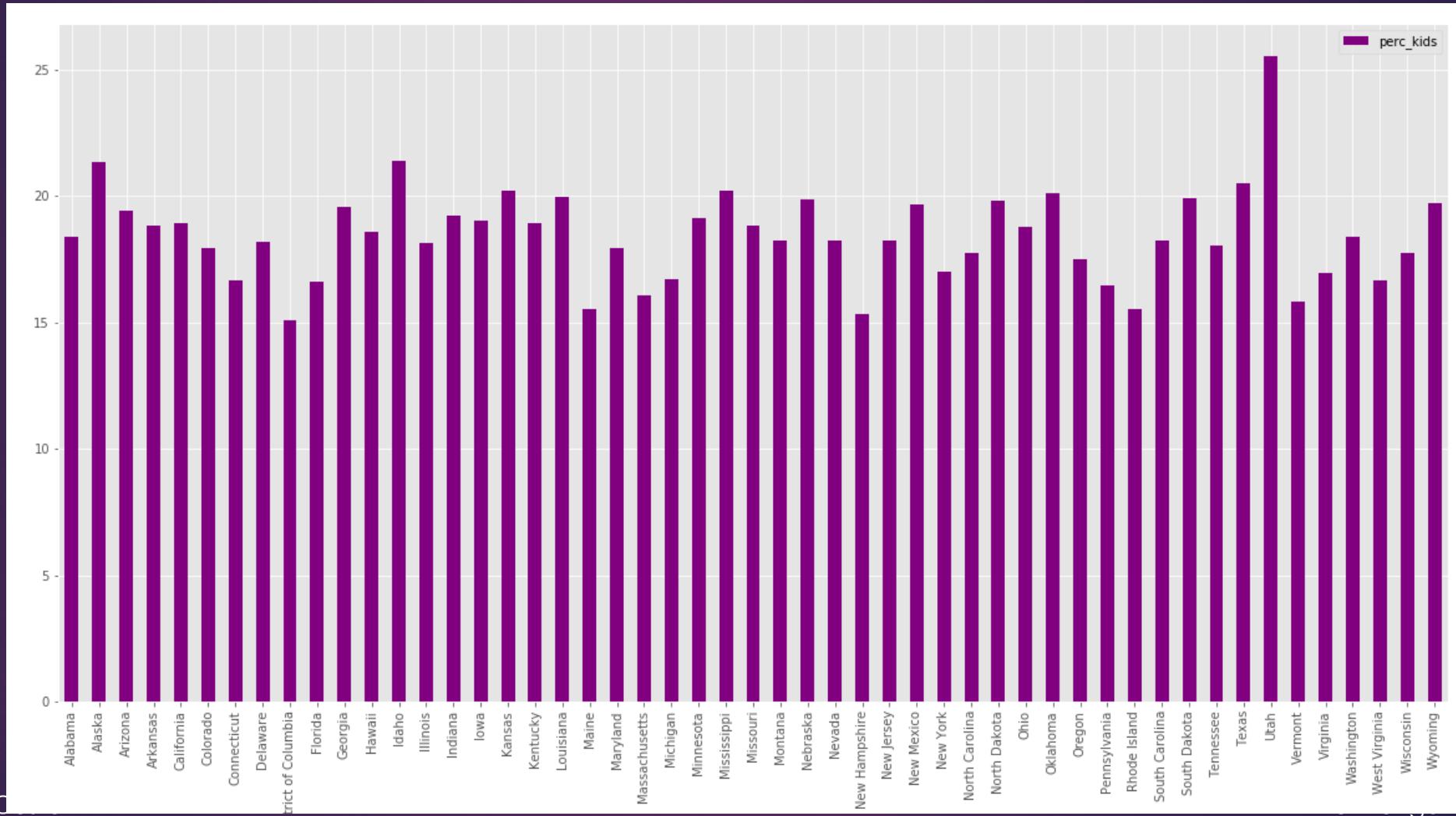
Percentage of working population



Percentage of men



Percentage of kids



Building Models

Selected predictors:

- Income per capita
- % of men
- % of worker
- % of people no insurance
- % of kids
- Number total population
- N° per person:
 - N° industries;
 - N° industries +20 employees;
 - Total employees;
 - Total paid to employees;
 - Production Hours;
 - Total paid production employees;
 - Amount paid shipping

Separating data for training and testing

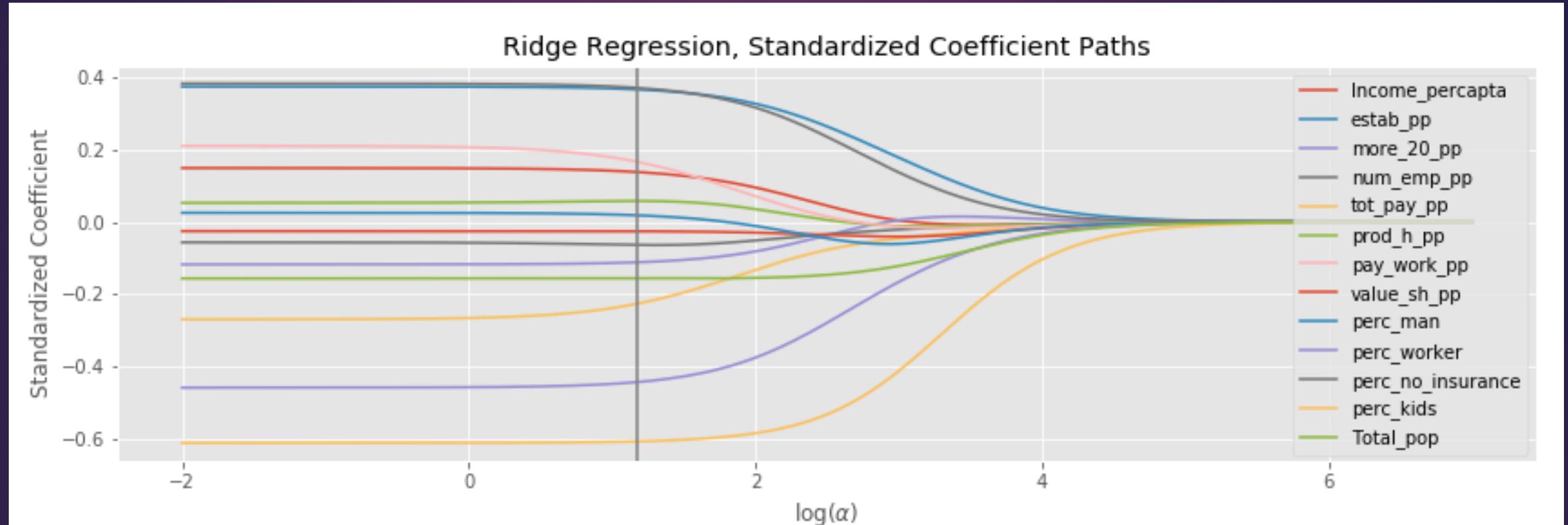
Training = 75%

Test = 25%

Ridge

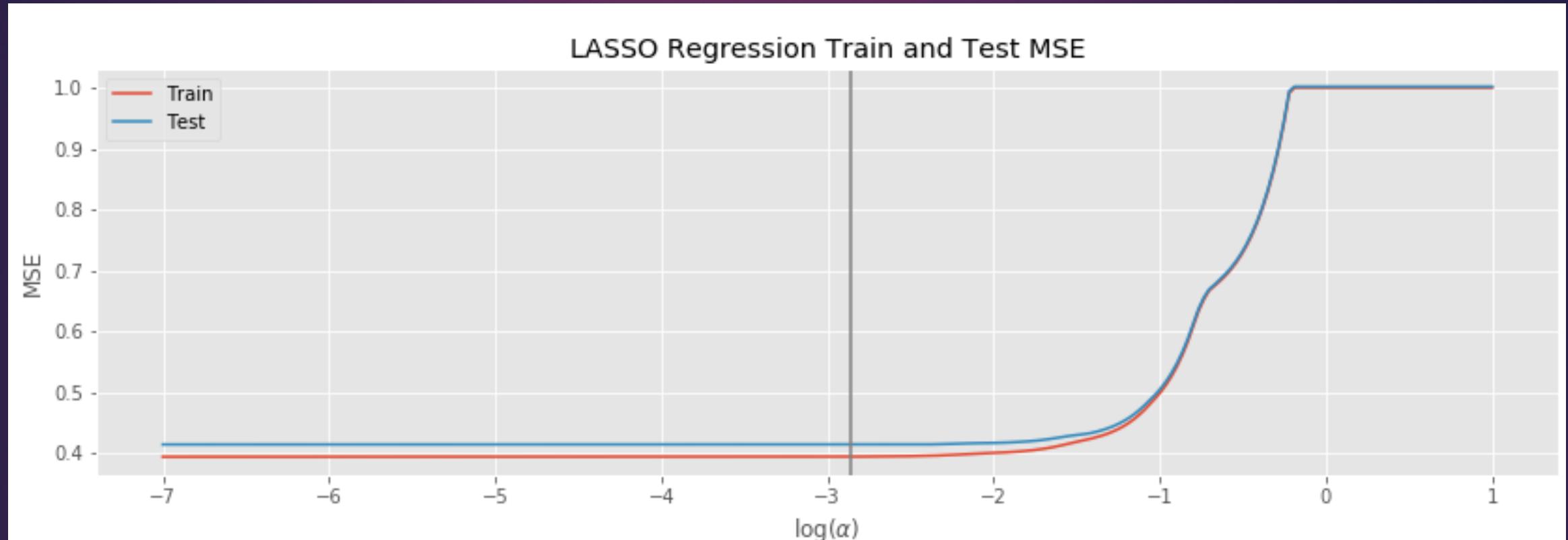


Ridge

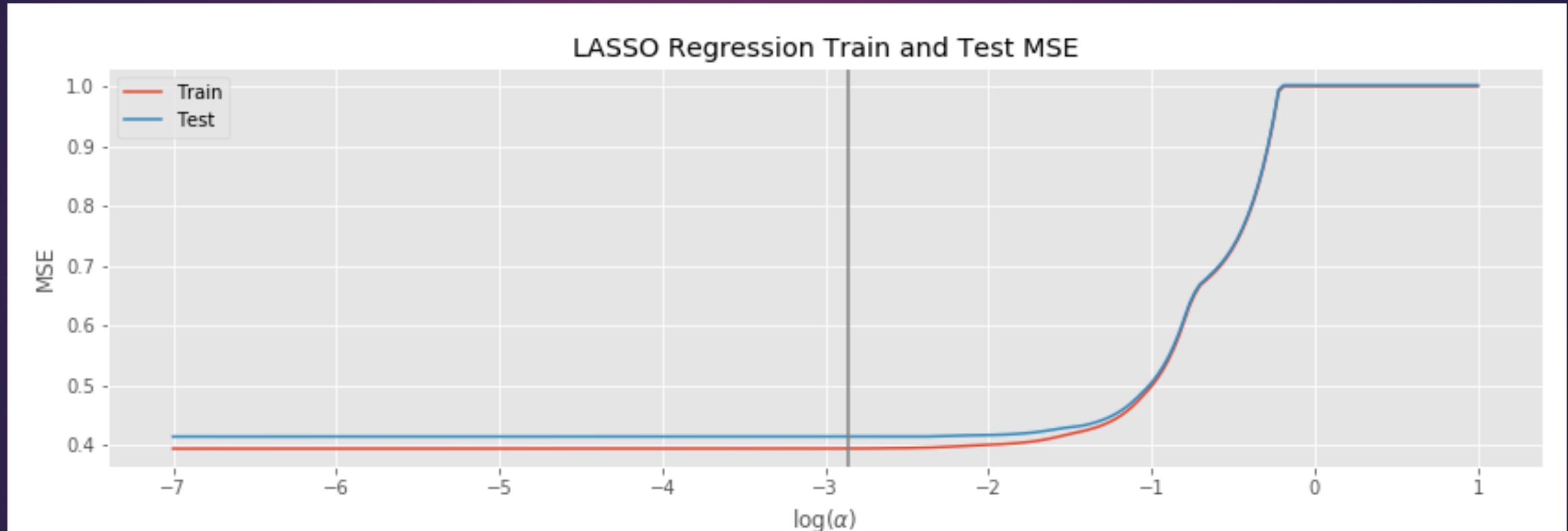


Log Ridge Optimal Alpha: 2.72

Lasso



Lasso



Log Lasso Optimal Alpha: -6.57

Linear Regression Coefficients

Name	Parameter Estimate
<hr/>	
Income_per capita	0.149
estab_pp	0.374
more_20_pp	-0.118
num_emp_pp	-0.056
tot_pay_pp	-0.269
prod_h_pp	0.053
pay_work_pp	0.210
value_sh_pp	-0.026
perc_man	0.025
perc_worker	-0.458
perc_no_insurance	0.383
perc_kids	-0.611
Total_pop	-0.156

Comparing Errors

Model	Train Part (75%)	Test Part (25%)	All Data
Ridge	0.395	0.445	0.400
Lasso	0.395	0.444	0.400
Linear Regression	0.395	0.444	0.400

* Very similar errors

Conclusion:

- To improve the model, it would be necessary to include other information that may influence the variation in the percentage participation of seniors in the municipalities.

Learning :

- Search the data directly from a source;
- Organization and selection of relevant information;
- Application and comparison of models (Ridge, Lasso, LR and CV (LR));
- Need to better analyze the information available before selecting a subject for work.

Frustration:

- Not being able to get more data related to my theme to create a more effective model.