An Apple A Day

TJ Barclay
Anissa Khan
Archana Ramakrishnan
Grant Gollier
Anna Fritz
John Quitno

Overview

Goal

Study the relationship between factors like income, food security, education performance, demographics

Problems

- Group counties based on the above factors
- Predict test scores based on food access and income
- Identify food desert counties

The industry

Education

- Adjusted test scores for grades 1-8

Food Security

- Percentage of people with supermarket access at various distances

Income

- Average income values

* all categories were analyzed per US county

Data Exploration

- Merging food desert and education datasets on County and State
- Aggregation method based on datatype
  - Counts/Totals = sum
  - Percentages = population weighted sum
  - Booleans = population weighted sum (percentage of population with that property)
  - Income = average of medians
- NaNs filled with column average

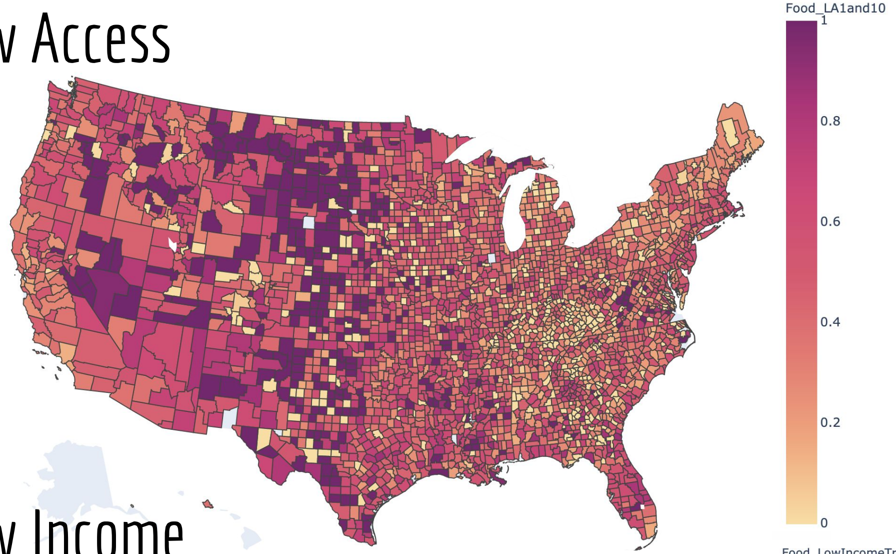
Clustering

Can we group counties based on educational performance, economic status, and food scarcity?

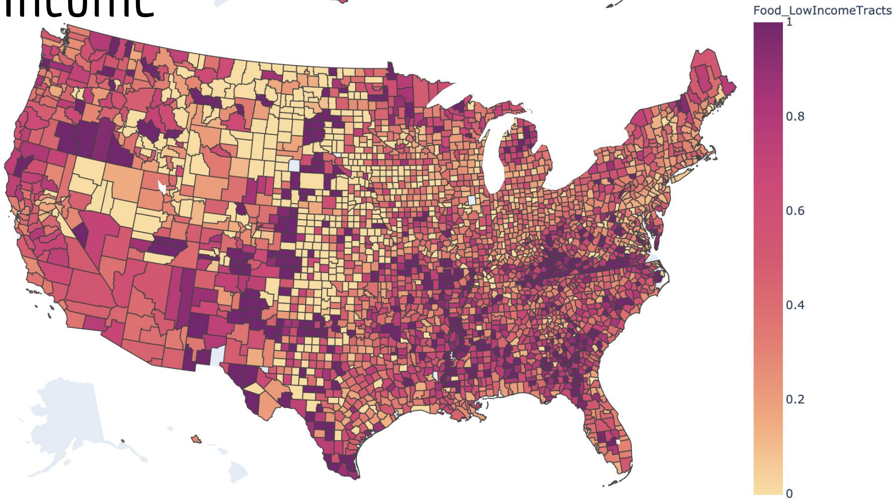
Data Used

- Low Food Access measurements
  - Food_LA1and10
  - Food_LAhalfand10
  - Food_LA1and20
- Low Income measurement
  - Food_LowIncomeTracts
- Educational Performance measurement
  - Educ_mn_avg_ol

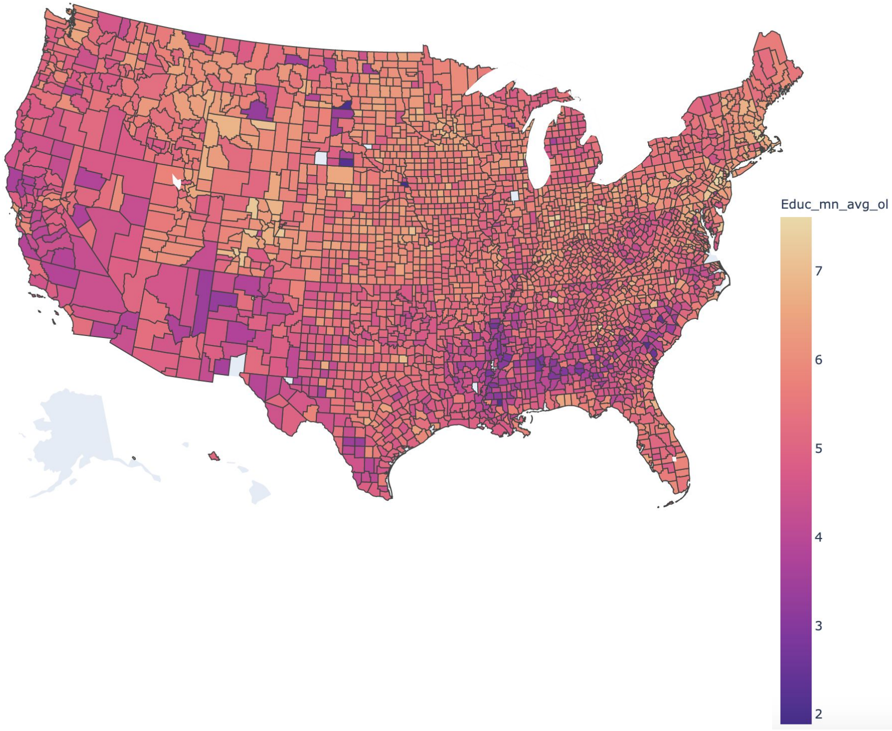
Low Access



Low Income



Educational Outcomes



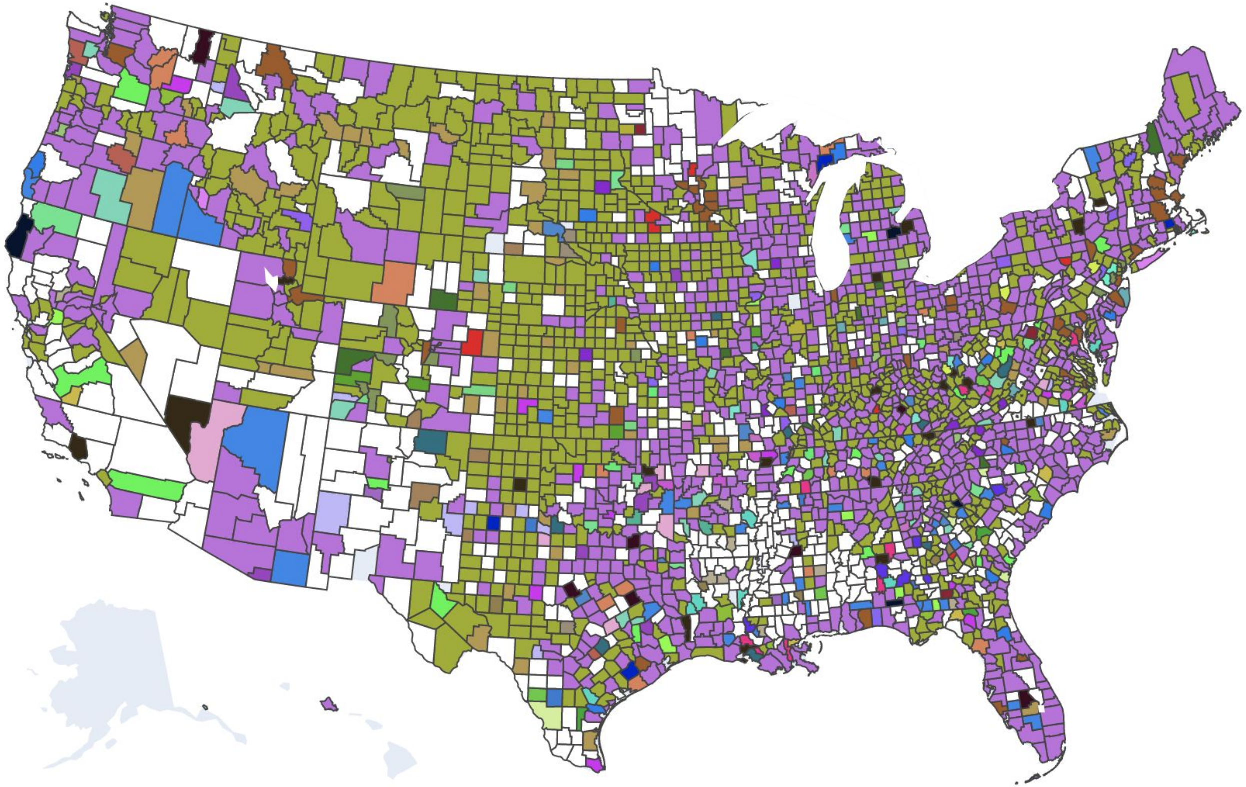
Algorithms

- DBSCAN
- Spectral Clustering
- Agglomerative

DBSCAN

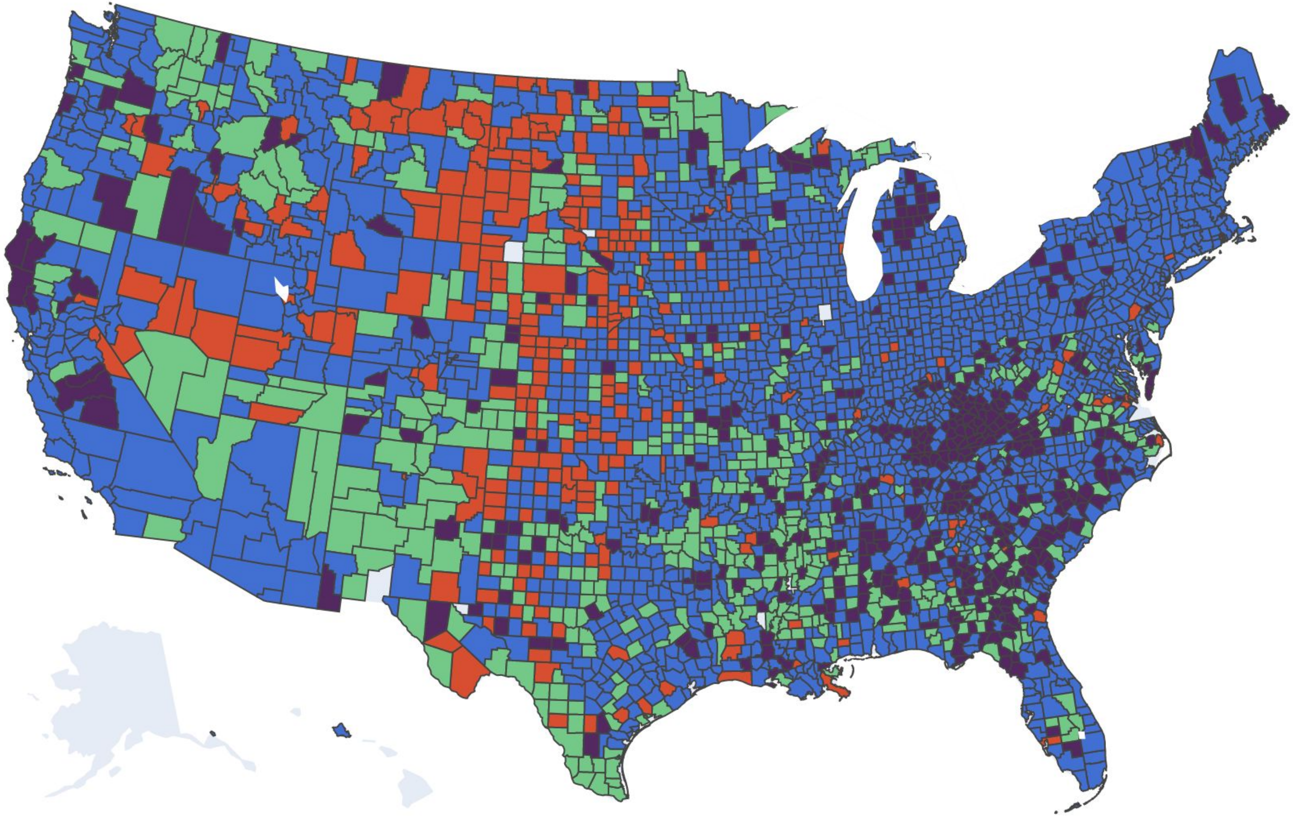
clusters: 57

noise points: 560



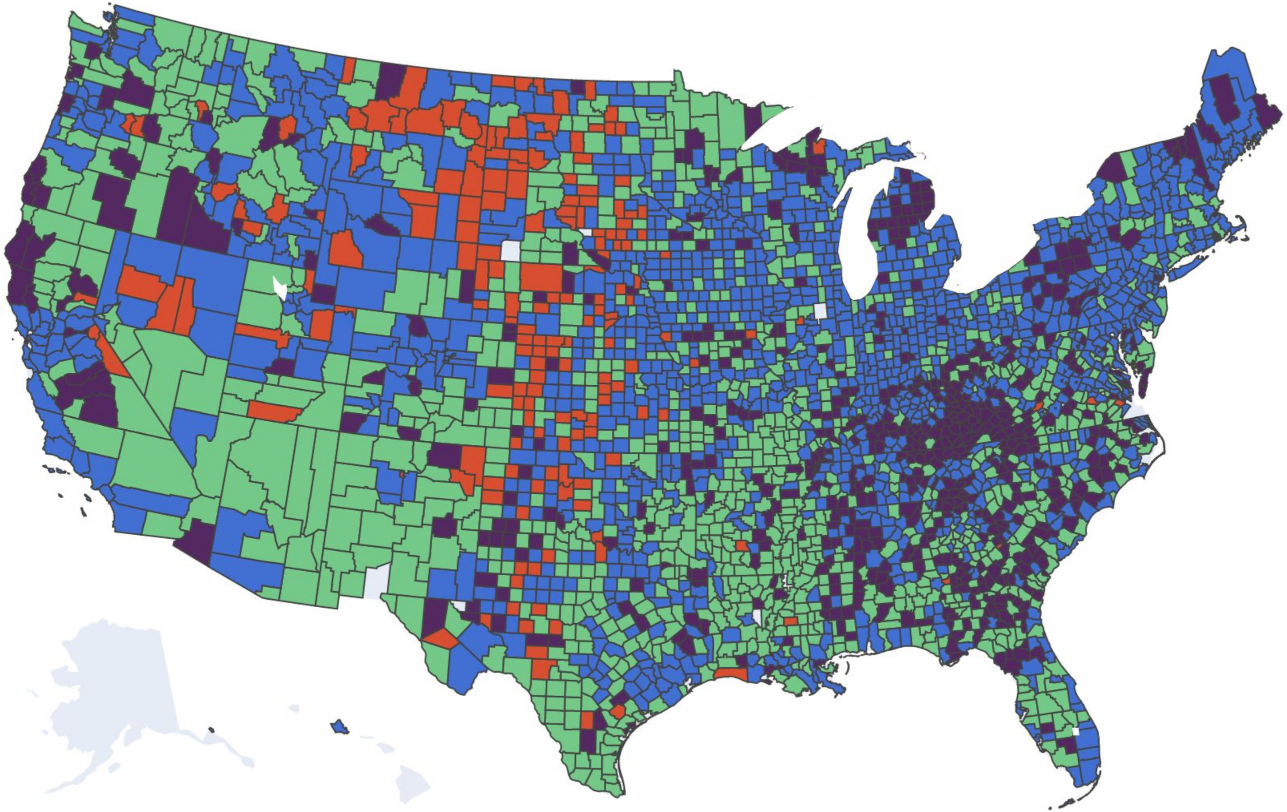
Spectral

clusters: 4



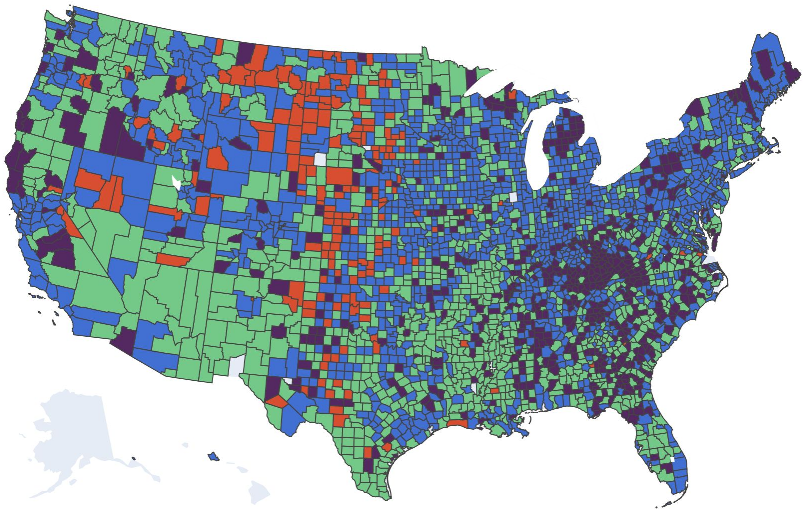
Agglomerative

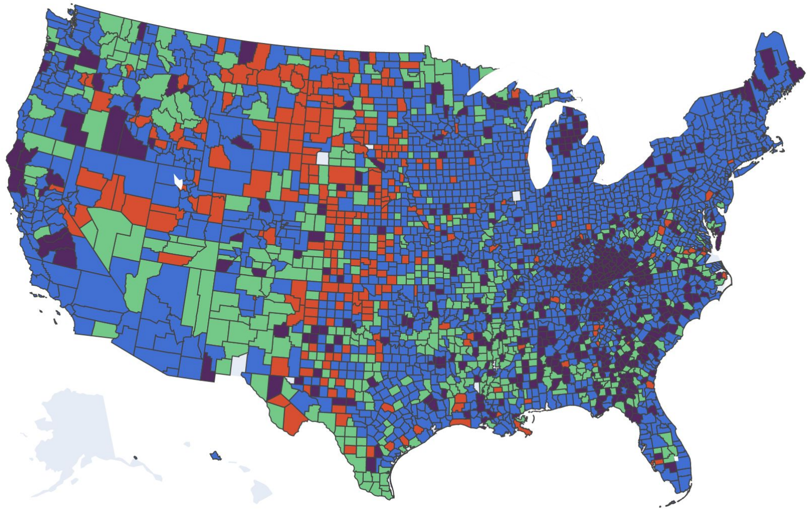
clusters: 4



Results

- Spectral and Agglomerative models produced the best results
- Red cluster (central)= reflects high income, low access
- Green cluster (primarily southern) = reflects low income/access/performance
- Purple cluster (TN/KY/VA region) = reflects low income/high access
- Blue cluster = less well defined





Classification

Can we identify if a county is a food desert or not?

Classification Problem

- **Goal:** identify if a county is a food desert
- **To predict**, we use:
  - Demographic information
  - Academic performance
  - Income etc.
- **To test** our results, we use:
  - Columns with food desert flag

Data Used

- Removed totals, senior, and location data
- Compare counties better with percentages and averages
- Allows to compare counties in a location agnostic manner
- Education data is focused on children

Classes Using Binning

- Binary values turned into percentages after tract aggregation
- Need some way to treat as classes
- Bin the values – we chose 20% increments
- Fairly arbitrary – definitions do exist for food desert, but are tract based

Principal Component Analysis (PCA)

- Dimensionality reduction
- Minimizes information loss
- Maximizes variance

How We Performed PCA

- Standardized the data
  - mean is 0
  - variance is 1

- Create PCA model
  - 95% number of components

- Fit PCA on the training set

- Apply the mapping

Models

- Logistic Regression
- Decision Tree Classifier

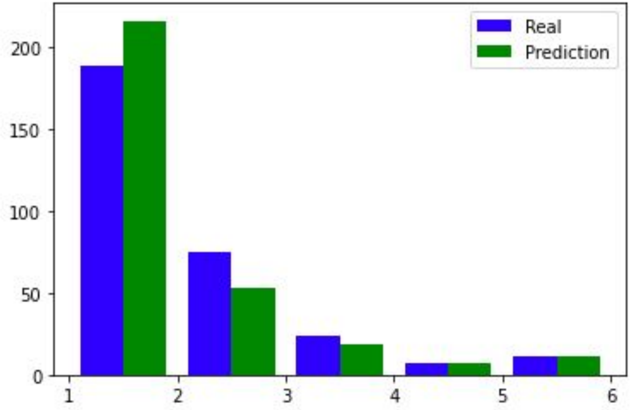
- Also tried
  - Neural Networks (MLPClassifier)
  - Random Forest

Logistic Regression Accuracy

- Accuracy: 72.86%
- PCA normalized data

```
Logistic Regression Classifier with  Food_LILATracts_1And10 as label:
________________________________________________________________________

Accuracy: 72.87582%
Correct: 223
Off by 1: 74
Off by 2: 9
Off by 3: 0
Off by 4: 0
```

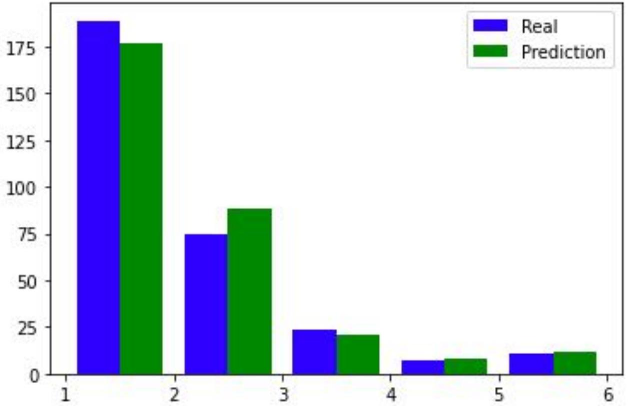


Decision Tree Accuracy

- Accuracy: 75.49%
- Original data

```
Decision Tree Classifier with  Food_LILATracts_1And10 as label:
_______________________________________________________________

Accuracy: 75.4902%
Correct: 231
Off by 1: 74
Off by 2: 1
Off by 3: 0
Off by 4: 0
```



Regression

Can we predict test scores based on food access and income?

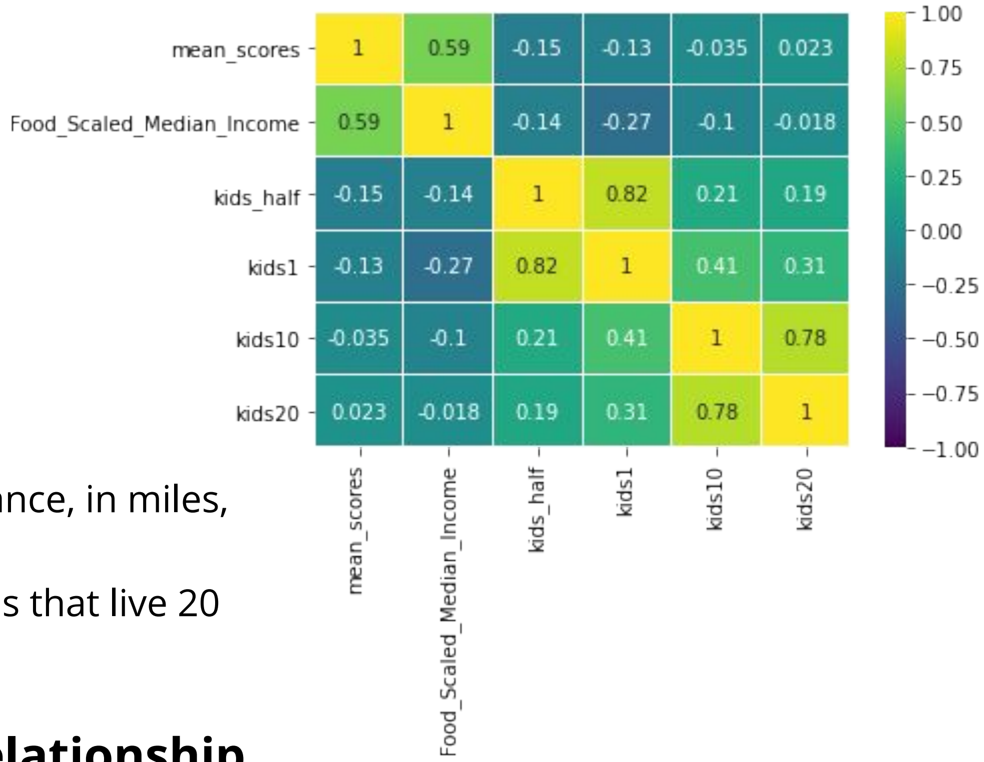
Data Used

Features Selected

- Mean Scores
  - What we will predict
- Median Income
- Kids_half, kids1, kids10, kids20
  - Percentage of kids that live x distance, in miles, from the nearest grocery store
  - Ex: kids20 is the percentage of kids that live 20 miles from the nearest grocery

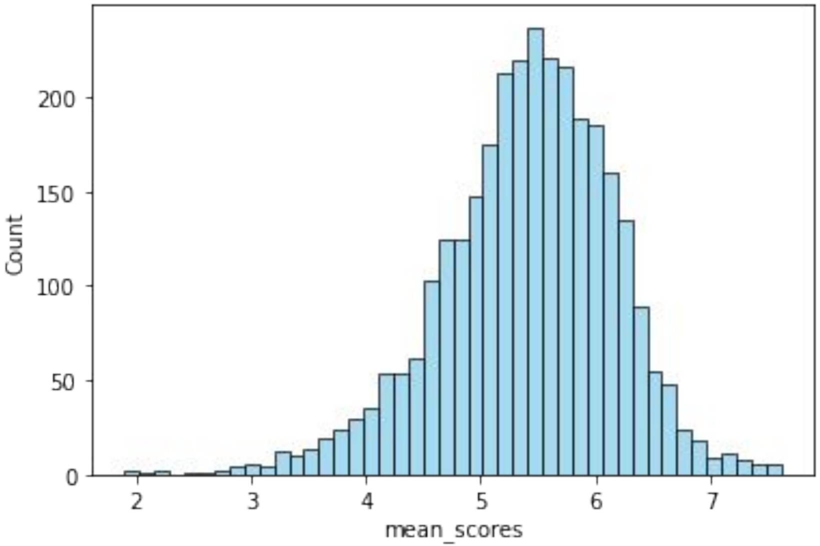
Using HeatMap to understand relationship

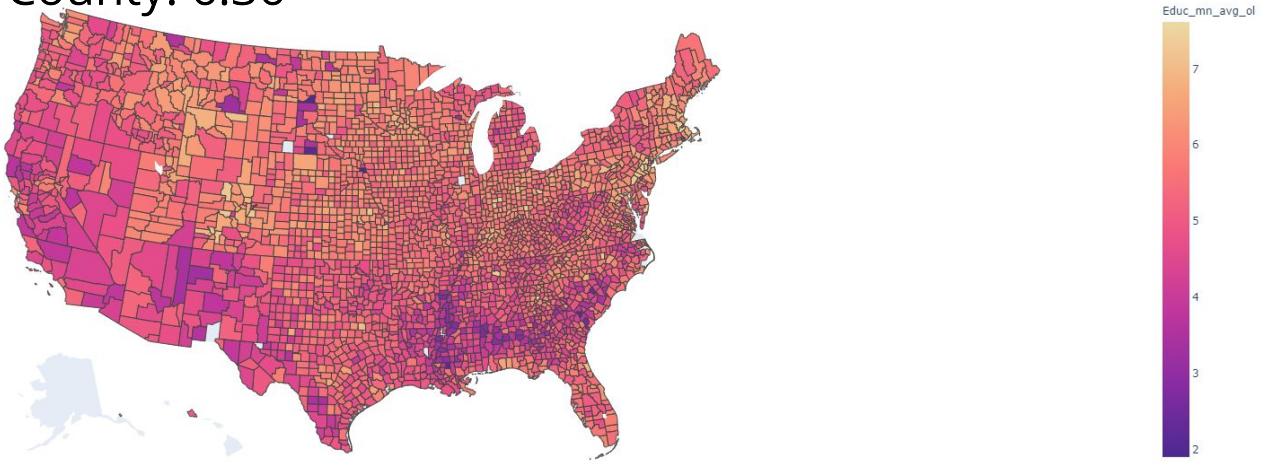
- Median income important
- Food access has some impact



Data Used (Continued)

- Fairly normal distribution:
  - Mean: 5.40
  - Median: 5.45
  - Min: 1.89
  - Max: 7.61
- Douglas County: 6.36



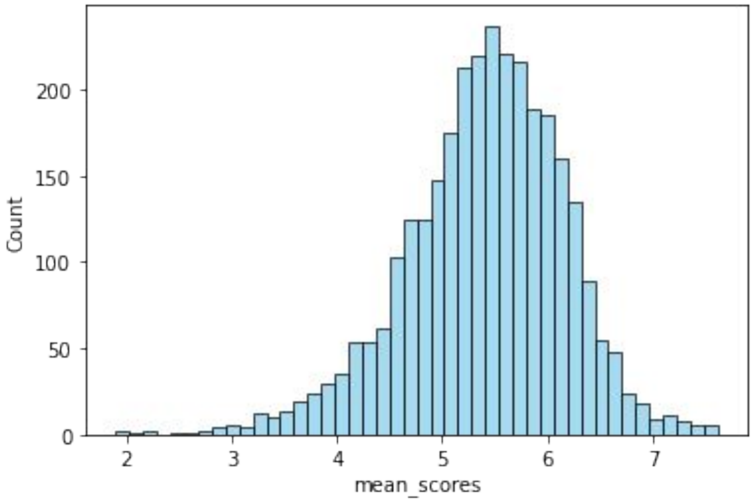


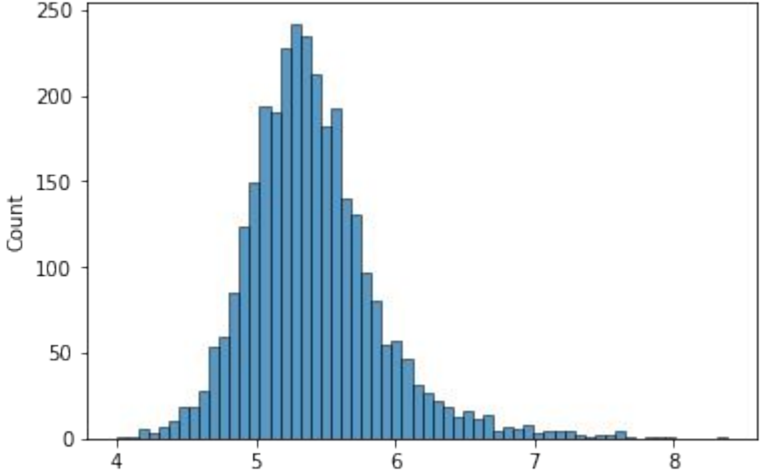
Overall Model Preparations

- Normalization
    - Poverty rate by converting to percentage in [0, 1]
    - Income levels with MinMaxScaler
- Train/Test Split
    - 10-Fold Cross Validation
- Douglas County Prediction

Linear Regression

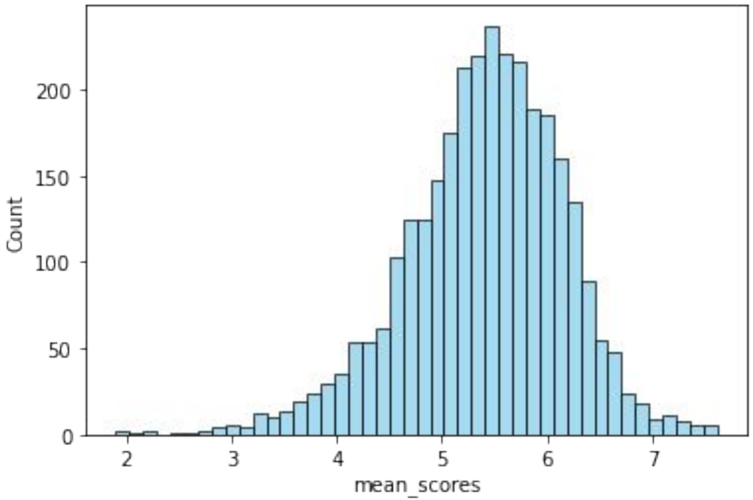
- Observations
  - Lower values being cut
  - Over prediction of higher values
- Results
  - Douglas County: 5.89
  - R^2: 0.25

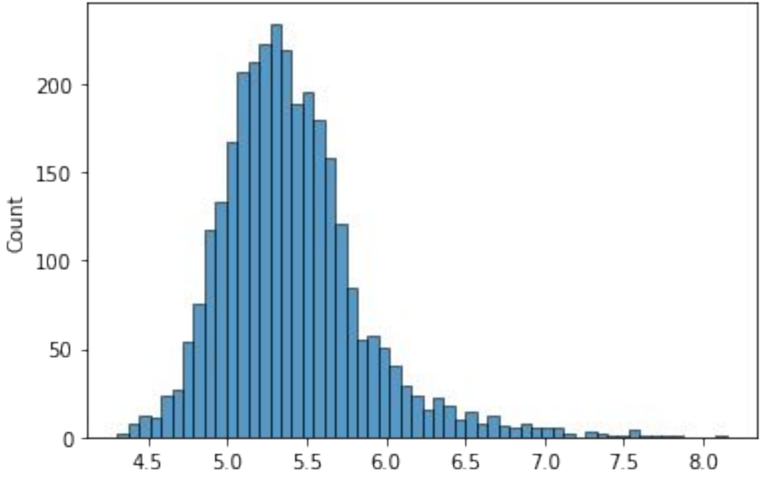




Ridge Regression

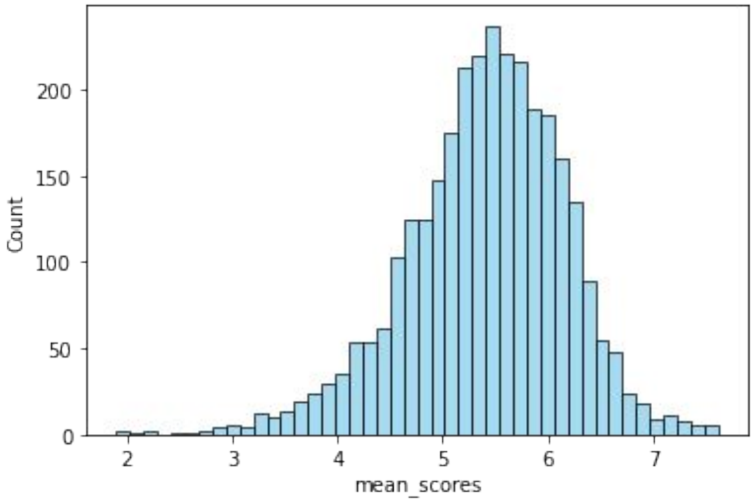
- Observations
  - Even more low scores being ignored
  - Similar over prediction of high scores
- Results
  - Douglas County: 5.85
  - R^2: 0.26

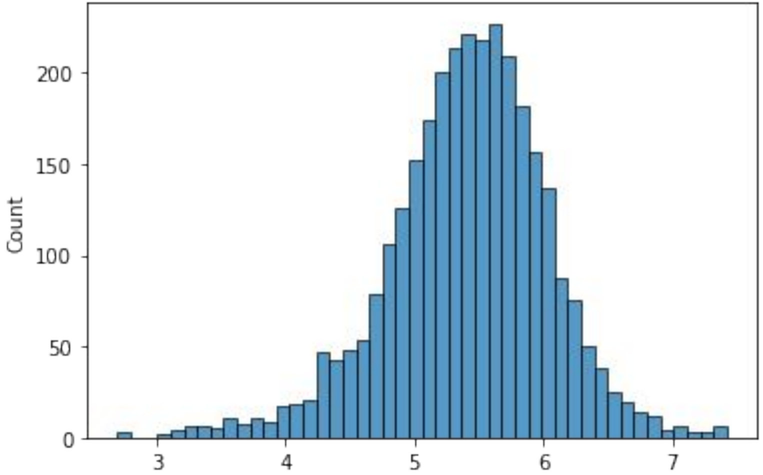




Random Forest Regression

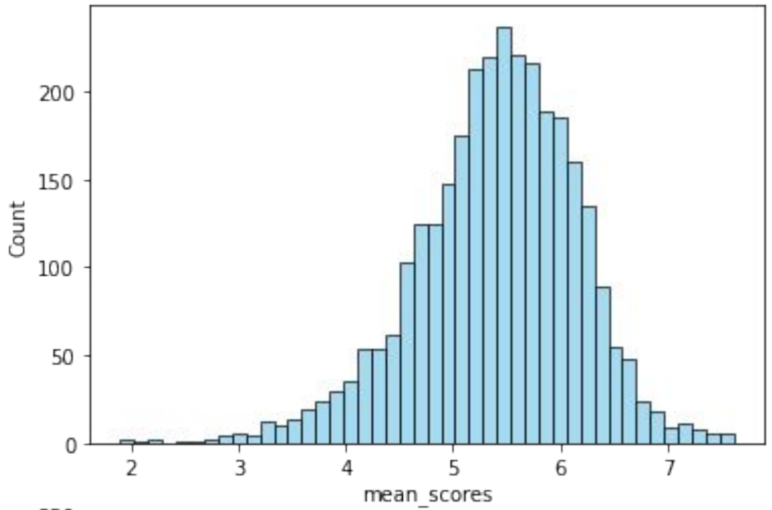
- Observations
  - Better coverage of lower scores
  - Better prediction of high scores
- Results
  - Douglas County: 6.10
  - R^2: 0.30

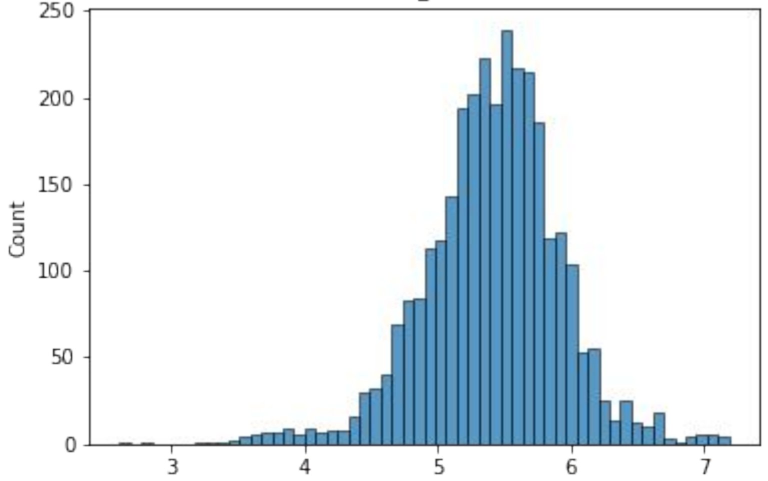




Gradient Boosting

- Observations
  - Under prediction of lower scores
  - Mediocre performance on higher scores
- Results
  - Douglas County: 5.93
  - R^2: 0.35





Conclusion

- **Classification**: we were able to identify if a county is a food desert or not with upto seventy five percent accuracy
- **Clustering**: we were able to cluster general regions together based on income, educational performance, and food access
- **Regression**: we were *somewhat* able to predict test scores based on food access and income