Evaluating and Predicting Food Assistance Based on Government SNAP WIC Data

Kevin Yin, John Sachenik, Kelley Shim, Aaron Chen

Our Team

Kevin Yin

Class of 2024

John Sachenik

Class of 2024

Kelley Shim

Class of 2024

Aaron Chen

Class of 2025

Agenda

1 Purpose

2 Data Exploration

3 Models

4 Insights

5 Conclusions

Purpose

Imagine you are in charge of allocating the appropriate amount of funds to a state-run food program...

you know the population demographics of the state, but how can you meet each different group's needs?

Key Questions

Which demographics are most important in predicting funding?

Which states are underperfoming compared to their peers?

How well can we predict funding based off demographics?

Purpose — Data — Models — Insights — Conclusions

Project Goals

01 Provide an accurate estimate of funding needed for an area based on certain demographics

02 Determine which demographics have the largest impact on funding

03 Visualize how state funding has changed over time and how different states are performing today

Data Exploration

Supplemental
Nutrition
Assistance
Program

Data Overview

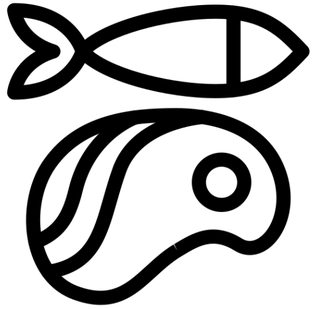
The data set was published on **Kaggle** by JohnM in 2019

Covers US Supplemental Nutrition Assistance Program for Women, Infants, and Children

Data ranges from 1969-2019

Provides state-by-state data for years 2012-2016

Data Insights



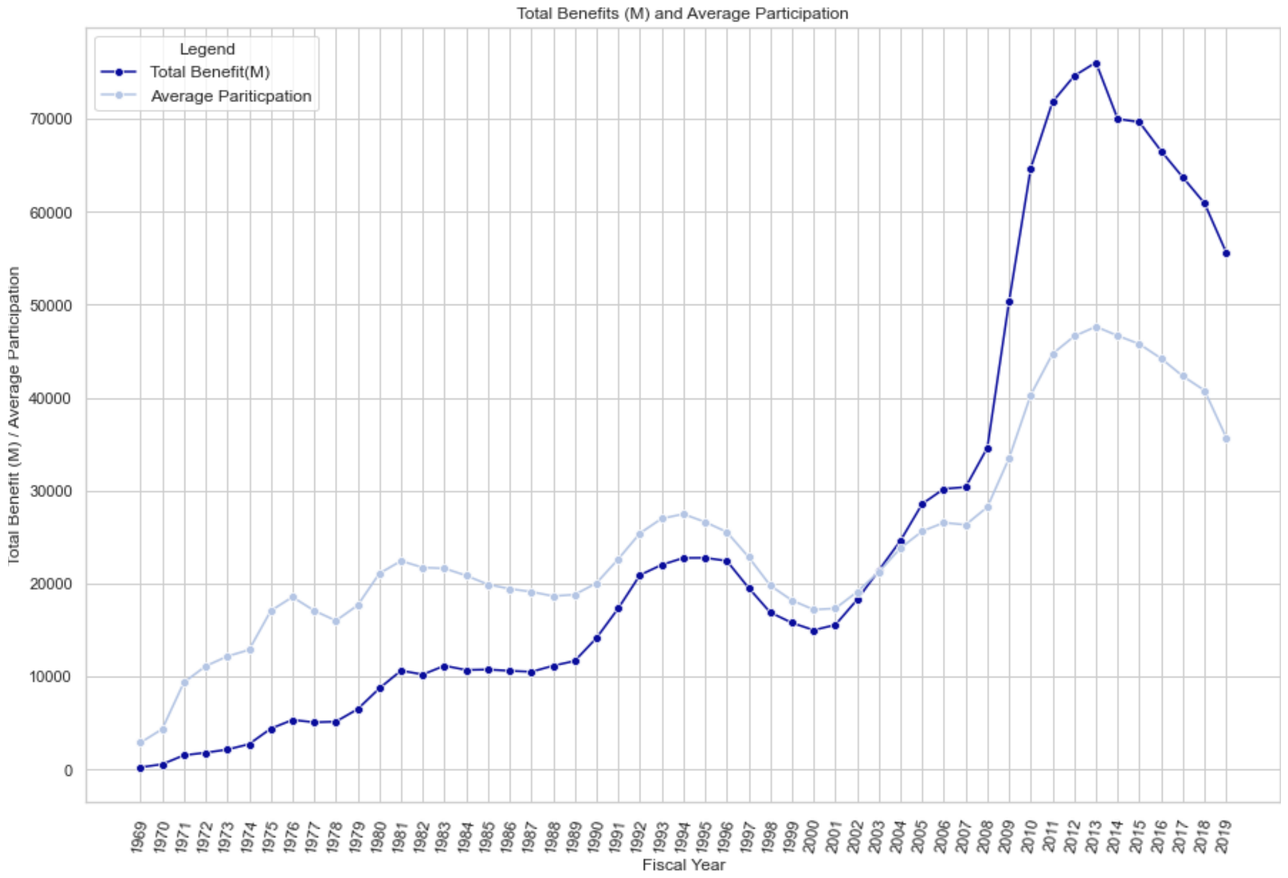
States with highest **Food Costs**:

1. California ($52,783,036.00)
2. Texas ($24,754,181.00)
3. New York ($23,474,332.0)

States with highest Population:

1. California (1,174,784)
2. Texas (858,671)
3. Florida (477,177)

Initial Data Visualizations



- Overall benefit and participation levels increased
- Starting from **2013**, benefit and particpation levels did decrease

Purpose — Data — Models — Insights — Conclusions

Population Distributions

Infants Fully Breastfed

Formula Fed Infants

Partially Breastfed Infants

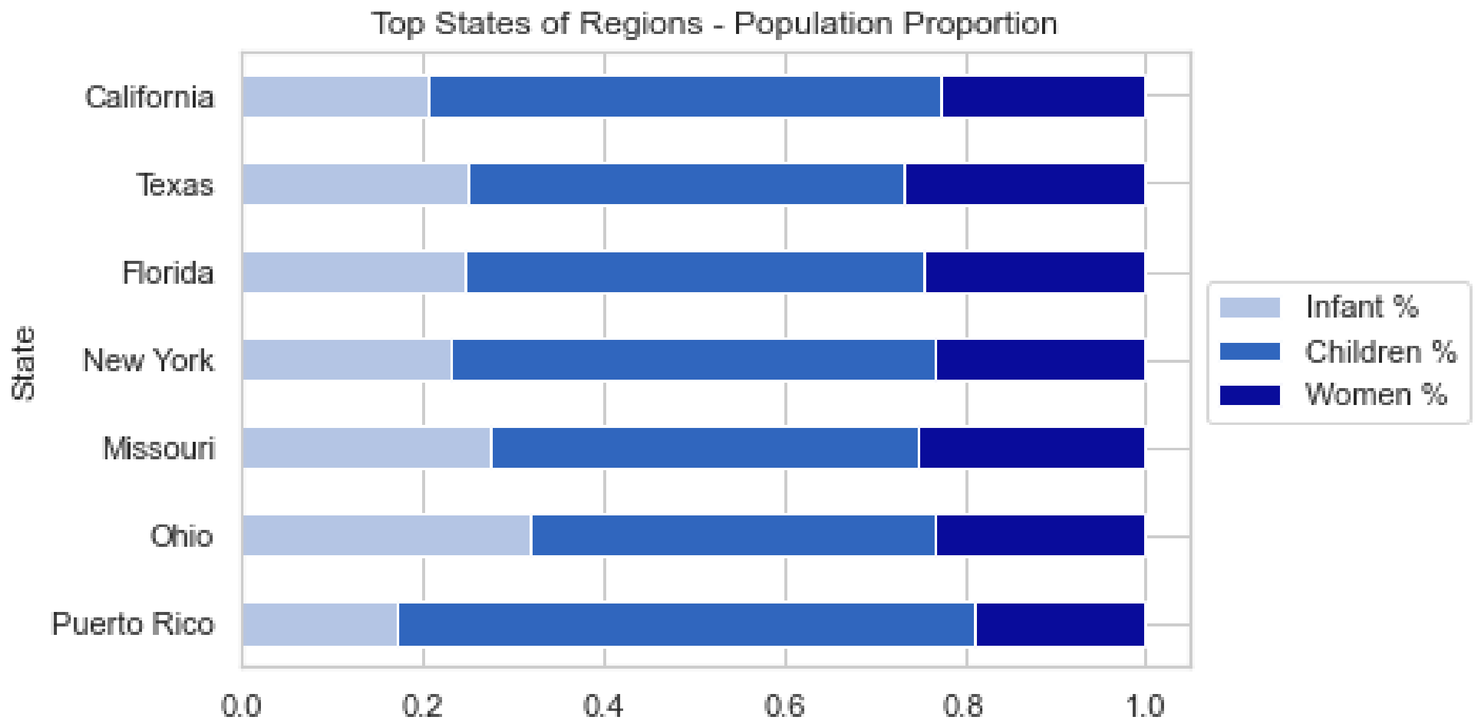
Children

Pregnant Women

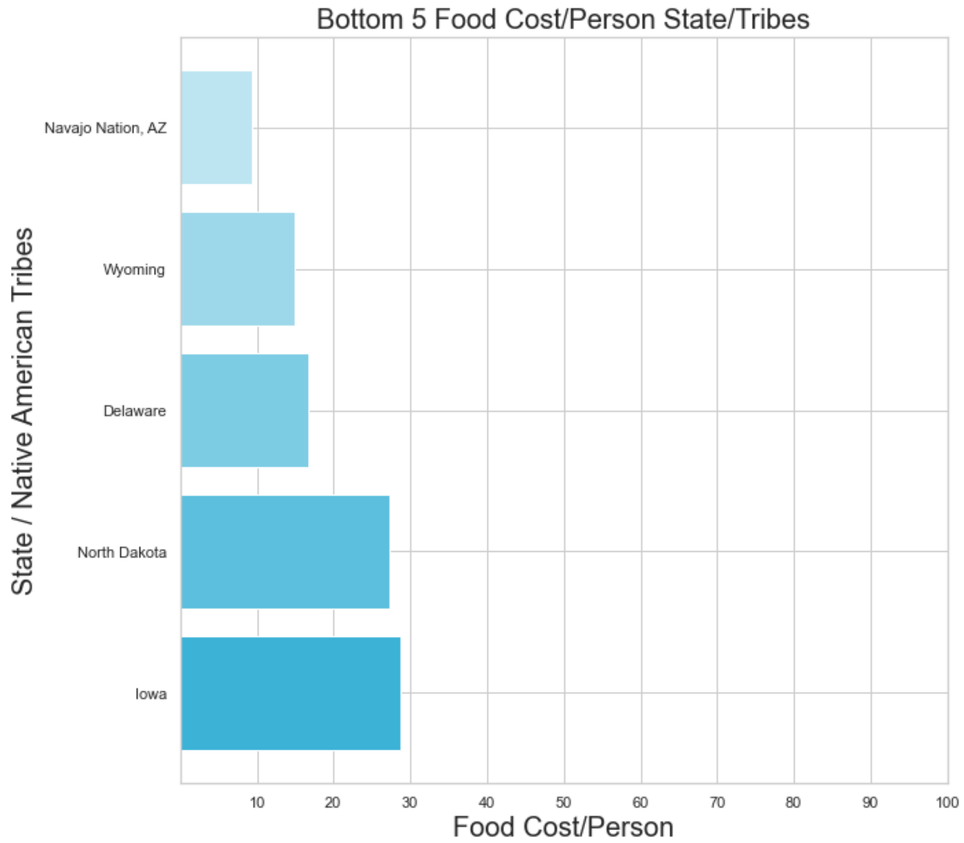
Postpartum Women

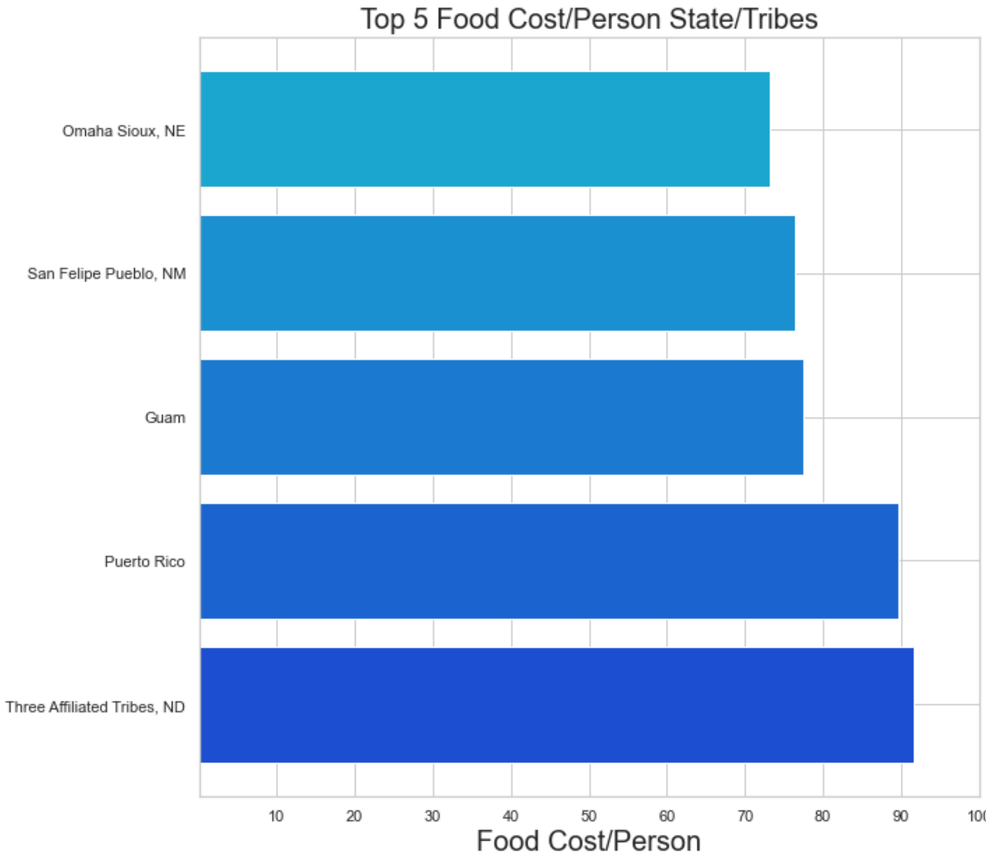
Initial Data Visualizations

With some variation, women and infants tend to make up 20% of the program each, while children make up about 60%



Top & Bottom 5 Cost/Population

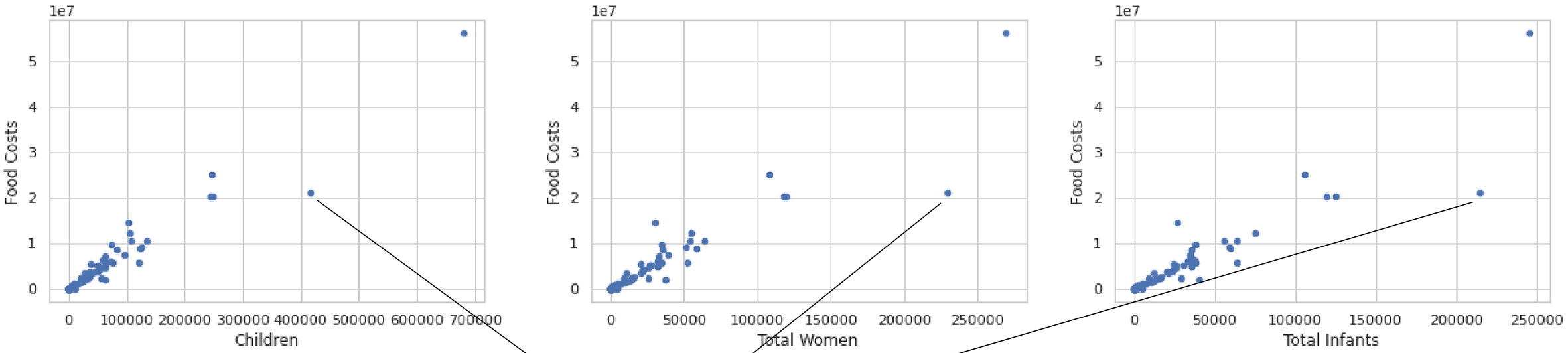




Purpose — Data — Models — Insights — Conclusions

Costs vs Population

As expected, we see positive trends across all groups when plotted with costs



Texas sticks out across all three demographics for recieving much less funding

Purpose — Data — Models — Insights — Conclusions

Further Data Visualizations

Political Affilitation?

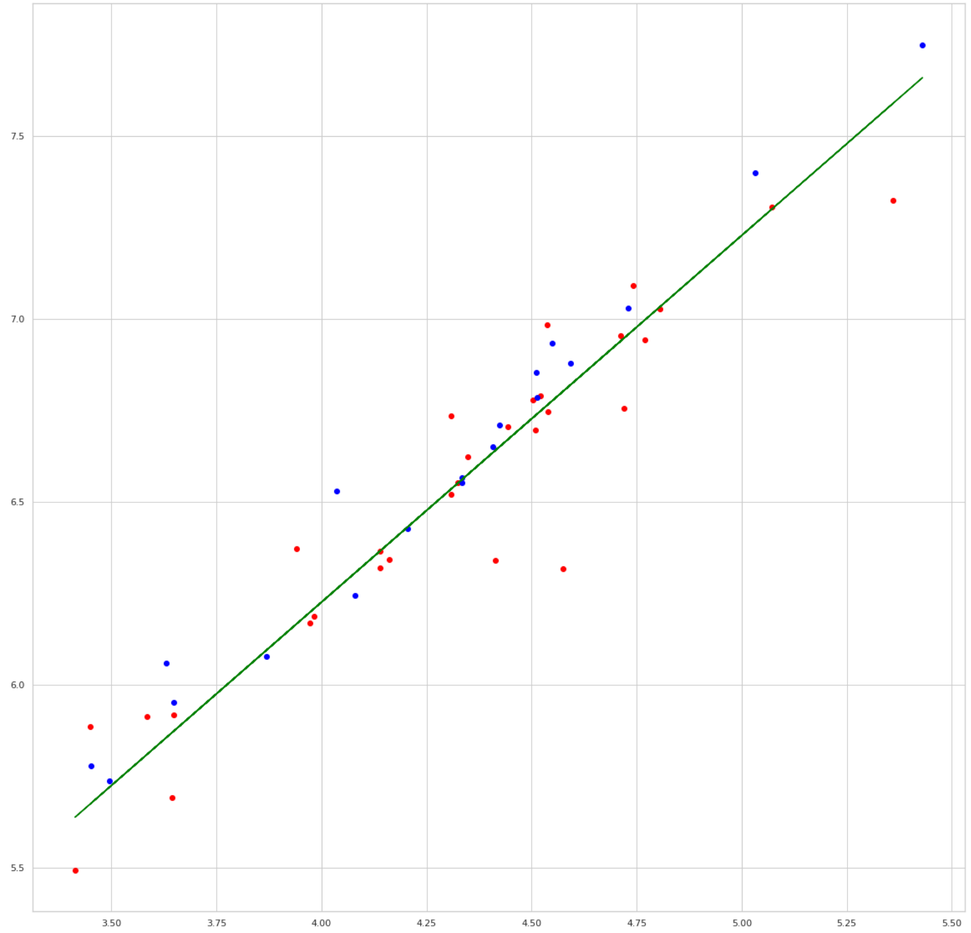
Red points = Republican States

Blue points = Democratic States

(based off 2016 election results)

Insights:

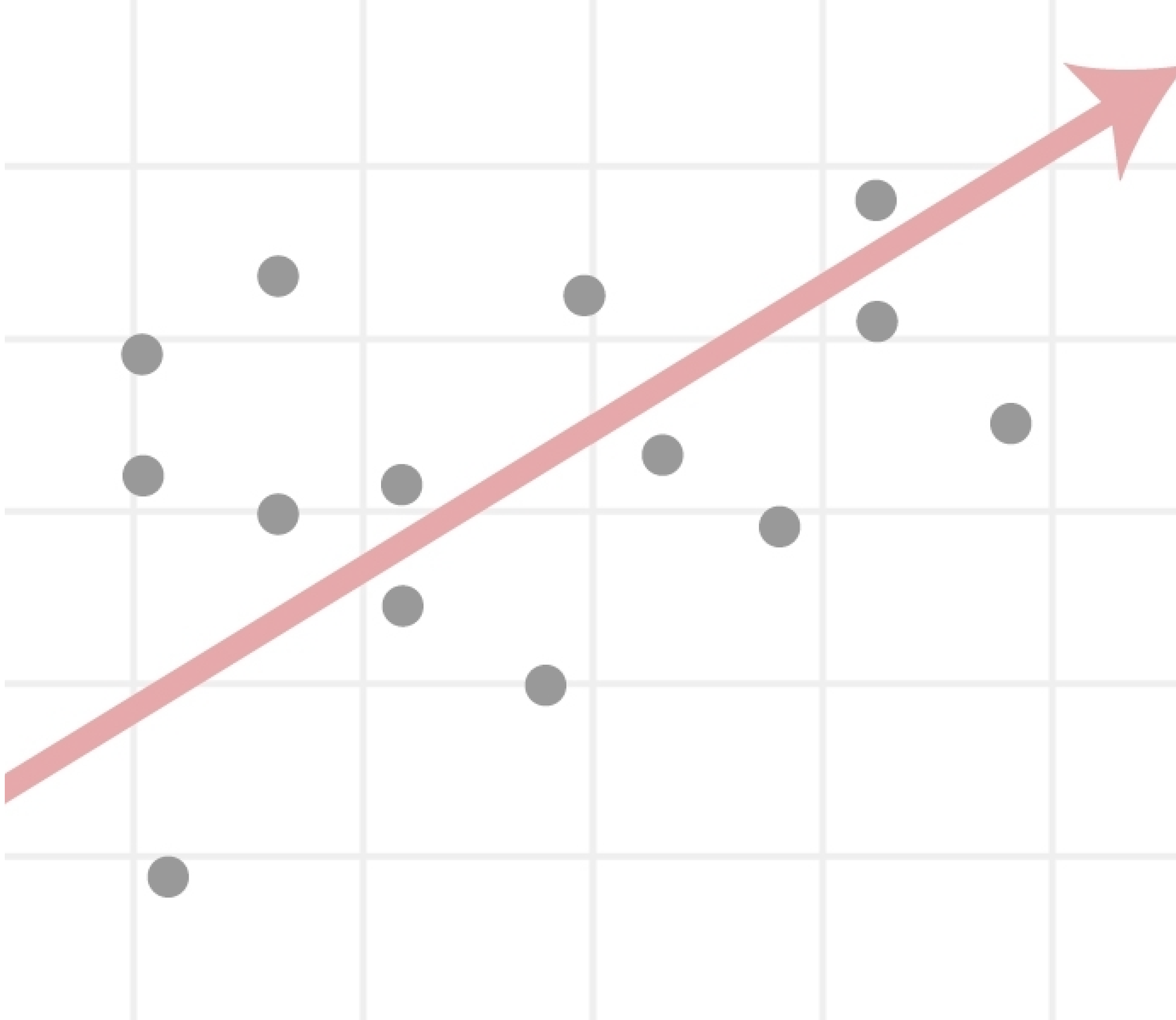
- 20% of democratic states fall below the regression
- 53% of red states fall below the regression



Statistical Significance

- Performed a 1 Tail 2 Proportion Z test
- Z Score of -2.33 which is significantly significant at 95%
- The proportion of states below the linear regression is **significantly different** between Democratic and Republican States

Models



Models Used

Linear Regression

Decision Tree Regression

Random Forest Regression

Why?

We are trying to predict a continous quantitative output based on our 6 demographic inputs

Linear Regression

Considers linear relationship of a dependent variable to one or more independent predictor variables

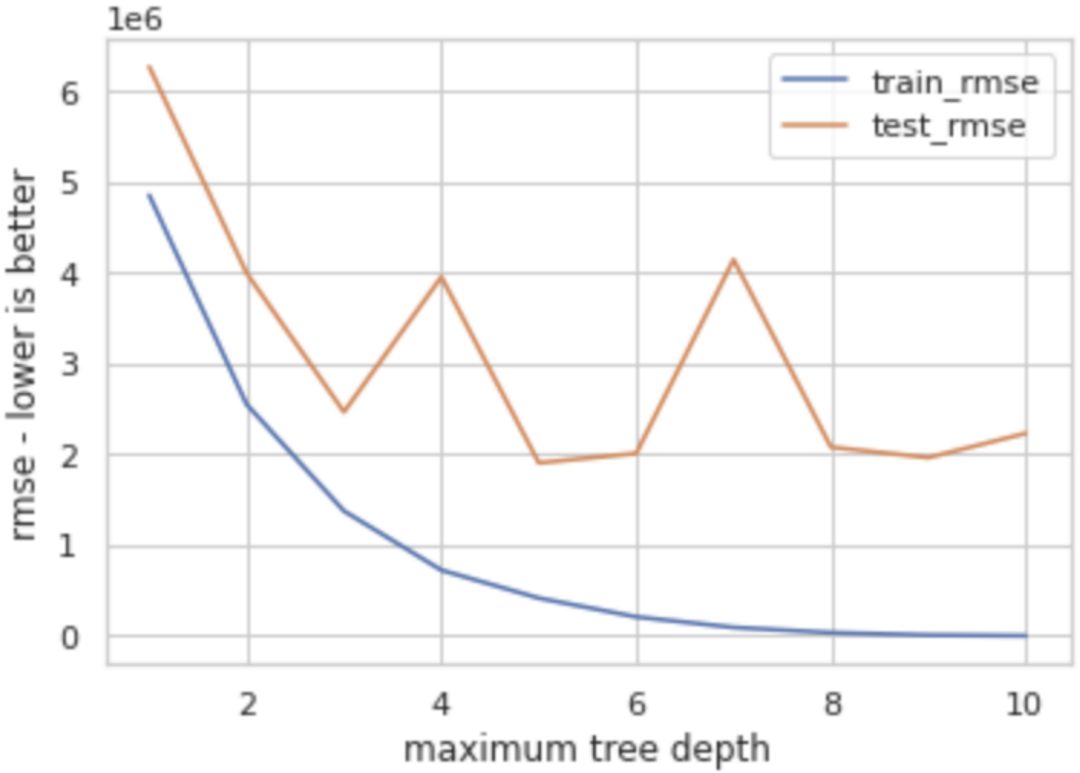
Decision Tree Regression

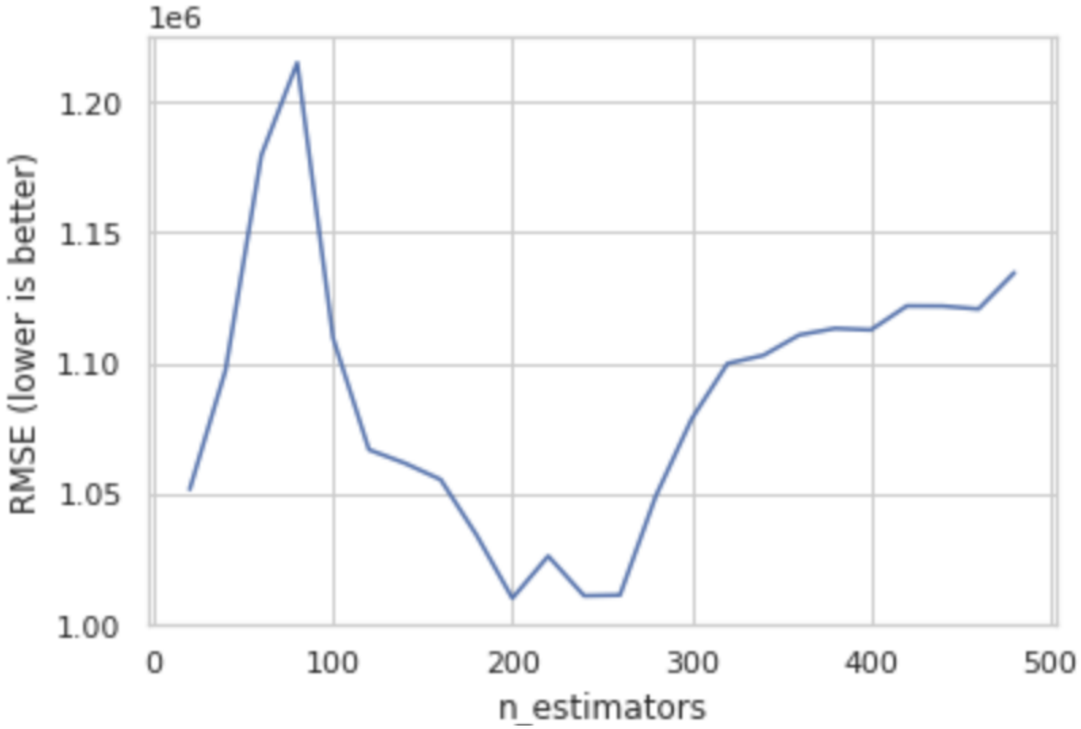
Makes decisions that relies on conditional control statements, increasing the homogeneity after each split.

Random Forest Regression

Constructs several decision trees training the model before outputting the mean of these classes.

Tuning Hyperparameters





Model Evaluation

	RMSE (2016)	RMSE (2015)	RMSE (2014)	RMSE (2013)
Linear Regression	1189	1989	1621	1303
Decision Tree Regression	1030	1852	1350	1134
Random Forest	960	1679	1156	890

Model Concerns and Limitations

1. Multicollinearity
2. Less influential features influencing regression
3. Decision tree sensitivity and limitations

Insights

Feature Importance

Calculated by Gini Feature Importance

1. Children
2. Pregnant Women
3. Formula Fed Infants

2013

0	Children	0.35
4	Breastfed Infants	0.18
3	Formula Fed Infants	0.15

2014

0	Children	0.28
1	Pregnant Women	0.27
3	Formula Fed Infants	0.14

2015

0	Children	0.22
1	Pregnant Women	0.21
4	Breastfed Infants	0.16

2016

0	Children	0.26
1	Pregnant Women	0.19
3	Formula Fed Infants	0.15

Model Insights

1. Children was consistently the most important feature across all years.
2. The top 3 features were not exactly the same throughout the years.
3. Random Forest was consistently the best performing model across years based on RMSE.

Concluding Insights

1

From our intial visualizations, we learned that:

1. Overall funding has been on the decline since 2012
2. Children make up the largest demographic of participants
3. Republican states are much more likely to be underfunded relative to democratic states

2

From our models, we learned that these demographics can be used to roughly predict state funding, but some error still remains

Conclusions: Limitations and Next Steps

Why not more accurate?

Shouldn’t funding be based on these exact demographics?

Our model does not consider:

Cost of Living

Political Affiliation/Size of local program

And much more

Next Steps

01 Consider cost of living for each state and regularize the funding according to these differences

02 See if factors like political affiliation can be brought into the model

03 Look into more timescale models and see how the overall decrease in funding effects each locality differently

Predicting Funding

Let’s say you have a state with the following demographics in 2016:

30,000 pregnant women

60,000 children

3,000 postpartum women

Our model predicts you will need to spend $10,180,435 on total food costs

As a Government Official...

You now know how to predict funding in your state!

Thank you!

Questions?

Appendices

Appendix A: Dataset

[91 rows x 10 columns]

	State	Food Costs	Children	Total Women	Total Infants	\
81	California	56217613.0	680664.0	269351.0	245469.0	
4	New York	25131432.0	246080.0	107622.0	105654.0	
40	Texas	21096330.0	415922.0	229206.0	214535.0	
20	Florida	20216722.0	244240.0	117747.0	119040.0	
77	Mountain Plains	20157097.0	249479.0	119275.0	124791.0	
..	...	...	...	...	...	
69	Santee Sioux, NE	9785.0	73.0	29.0	32.0	
9	Seneca Nation, NY	7131.0	83.0	47.0	61.0	
7	Indian Township, ME	4105.0	37.0	7.0	12.0	
8	Pleasant Point, ME	3437.0	28.0	8.0	14.0	
1	Maine	-121621.0	11477.0	4589.0	4940.0	

	Formula Fed Infants	Breastfed Infants	Partially Fed Infants	\
81	148203.0	51819.0	45447.0	
4	60510.0	9861.0	35283.0	

Appendix B: Linear Regression Code

```
# create X and y except now with more columns in X
feature_list = ['Children', 'Formula Fed Infants', 'Breastfed Infants','Partially Fed
'Postpartum Women']
X_mult = var2016[feature_list]
y_mult = (var2016['Food Costs'])

# instantiate and fit like last time
mult_linreg = LinearRegression()
mult_linreg.fit(X_mult, y_mult)

# print intercept
print("The y intercept:", mult_linreg.intercept_)

# pair the feature names with the coefficients
print(list(zip(feature_list, mult_linreg.coef_)))

# Evaluate R^2
y_mult_pred = mult_linreg.predict(X_mult)
print("R^2: ", metrics.r2_score(y_mult, y_mult_pred))

# Evaluate MSE
print("MSE: ", metrics.mean_squared_error(y_mult, y_mult_pred))

# Evaluate RMSE
print("Test set RMSE:",np.sqrt(metrics.mean_squared_error(y_mult, y_mult_pred)))
```

Appendix C: Decision Tree Code

[5]

```
X = var2016[feature_col]
y = var2016['Food Costs']

train_rmse,test_rmse = [],[]
for depth in range(1,11):
    dt = DecisionTreeRegressor(max_depth=depth)
    X_train,X_test,y_train,y_test = train_test_split(X,y,test_size=0.2,
                                                     random_state=20)
    dt.fit(X_train,y_train)
    train_rmse.append(np.sqrt(mean_squared_error(y_train,
                                                 dt.predict(X_train))))
    test_rmse.append(np.sqrt(mean_squared_error(y_test,
                                                dt.predict(X_test))))

sns.mpl.pyplot.plot(range(1,11),train_rmse,label='train_rmse')
sns.mpl.pyplot.plot(range(1,11),test_rmse,label='test_rmse')
sns.mpl.pyplot.xlabel("maximum tree depth")
sns.mpl.pyplot.ylabel("rmse - lower is better")
sns.mpl.pyplot.legend()
decision_tree = DecisionTreeRegressor(max_depth=3)
decision_tree.fit(X_train,y_train)
print("Decision Tree RMSE:",np.sqrt(mean_squared_error(y_test,decision_tree.predict(X.
```

Appendix D: Random Forest Code

```
X = var2016[feature_col]
y = var2016['Food Costs']

rf = RandomForestRegressor(n_estimators=500, bootstrap=True, oob_score=True, random_s
rf.fit(X_train, y_train)
y_pred_rf = rf.predict(X_test)

print("Random Forest RMSE:",np.sqrt(mean_squared_error(y_test,y_pred_rf)))
```