Food Insecurity in America

Help in a SNAP

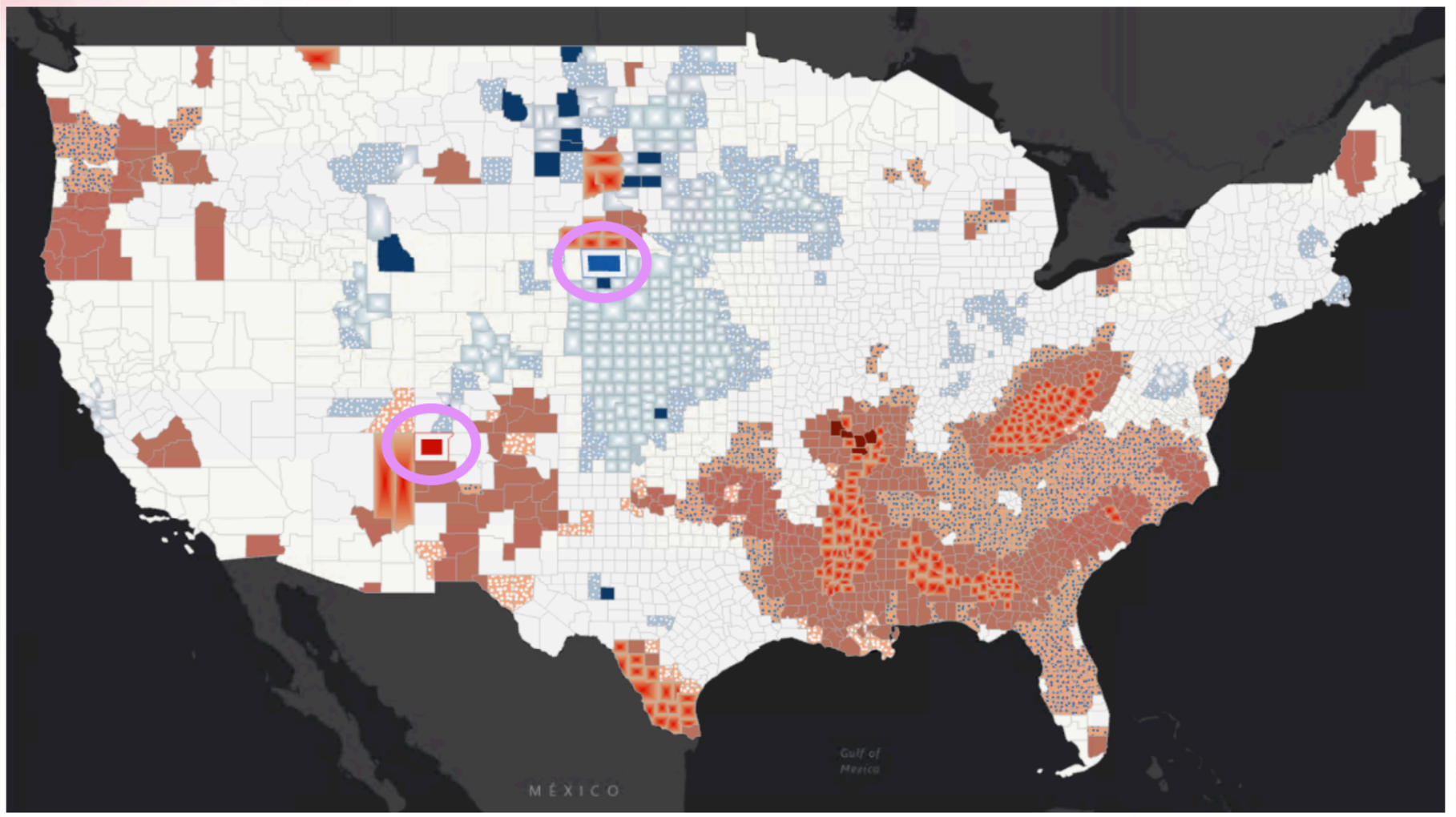
Which communities are at risk for food insecurity during COVID? I use a predictive model to determine areas at risk.

What do food insecurity community characteristics look like?

- I got data from the USDA SNAP program consisting of over 400k records of 800 features. Each record was an application
- I wanted to use a 10 year gap analysis by comparing 2007 and 2017 in my final predictive model.
- My target was a field called “CAT_ELIG” meaning out of all the applications received:
  - 1 = Eligible for benefits
  - 0 = Not eligible for benefits.

Narrowing down the data: GIS

- These two purple circles represent areas that are emerging hot and cold spots of SNAP dependency found from an ArcGIS MOOC on Spatial Analysis.
- The “hot spot” is San Juan County, New Mexico.
- The “cold spot” is Cherry County, Nebraska.
- **This narrowed the focus to Nebraska and New Mexico in 2007 and 2017 for the extremes of SNAP characteristics.**

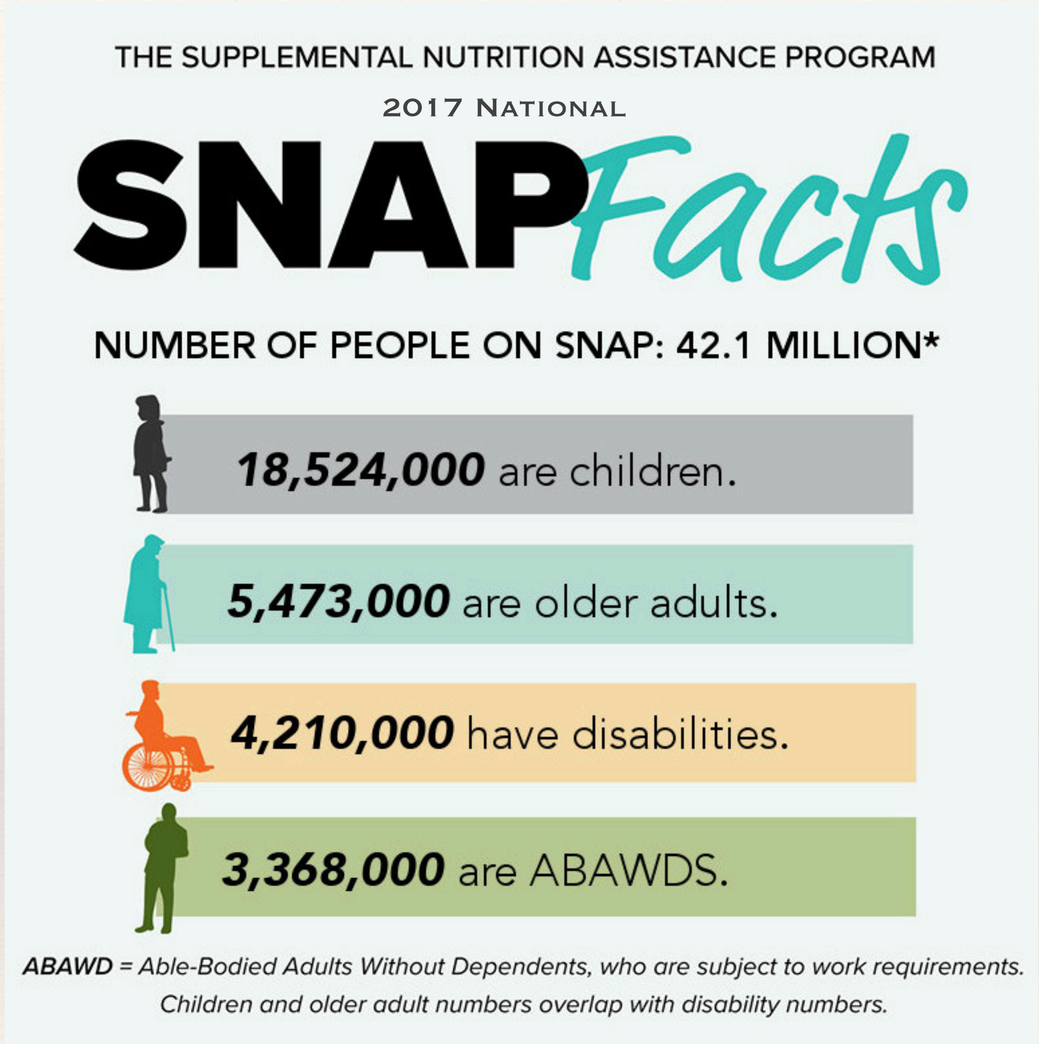


Application Counts in QC* Data

*QC data means only "complete" applications

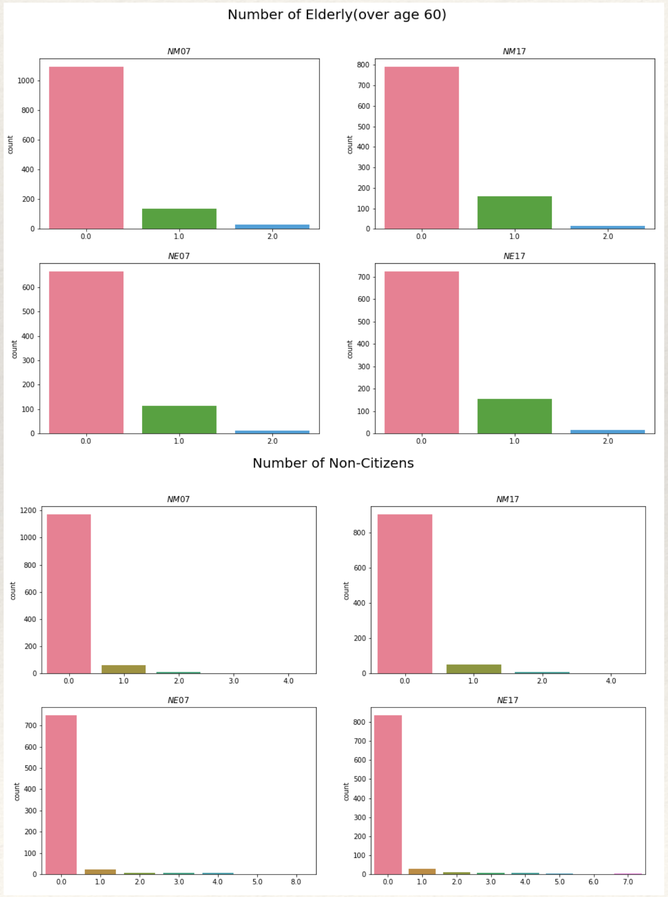
2007	2017
All: 47k	All: 45k
New Mexico: 1255 Nebraska: 791	New Mexico: 964 Nebraska: 894

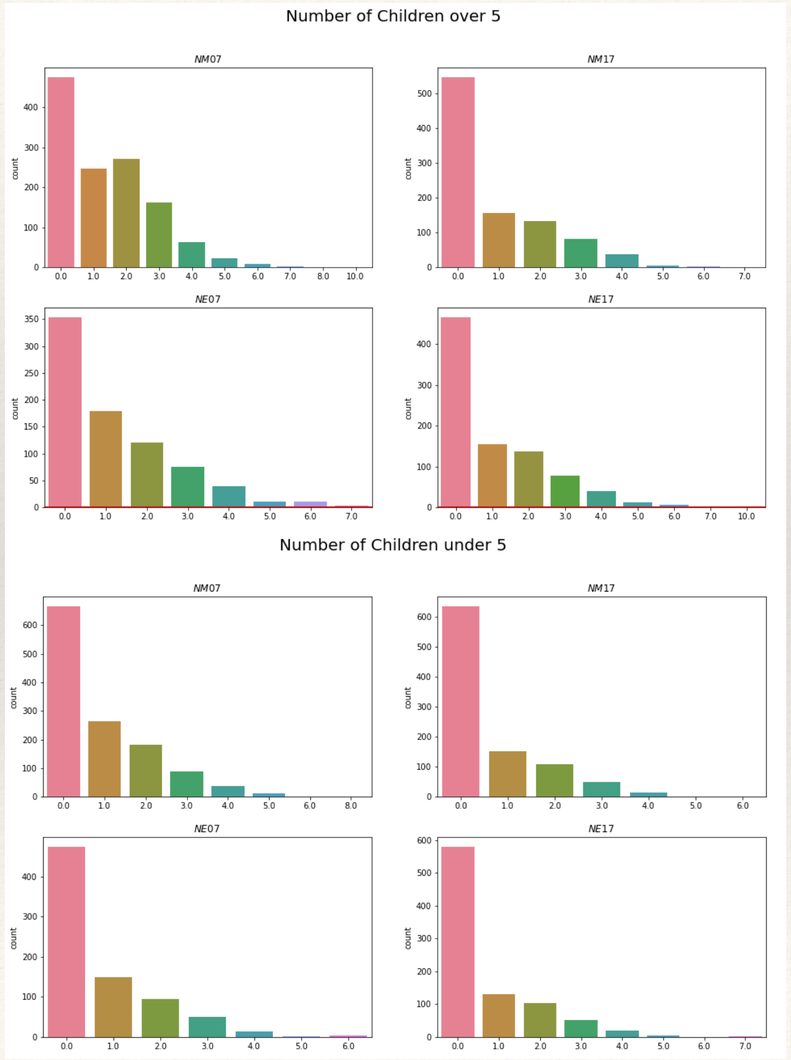
- New Mexico is vulnerable to wild swings in the economy, whereas Nebraska stays pretty consistent.

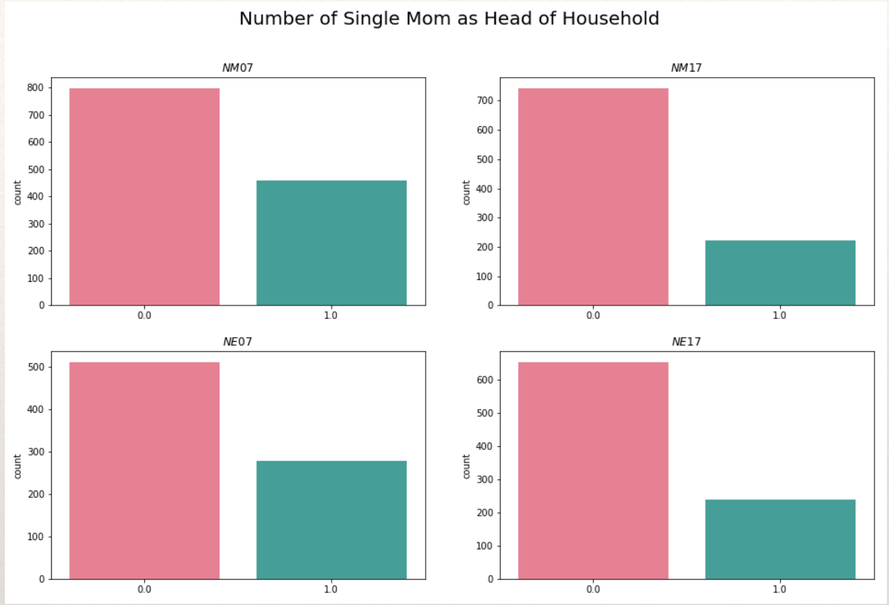


Initial Snapshot of the Data

- Most initial data shows no real trends of who applies for SNAP benefits.
  - Note: this is before nulls are removed.



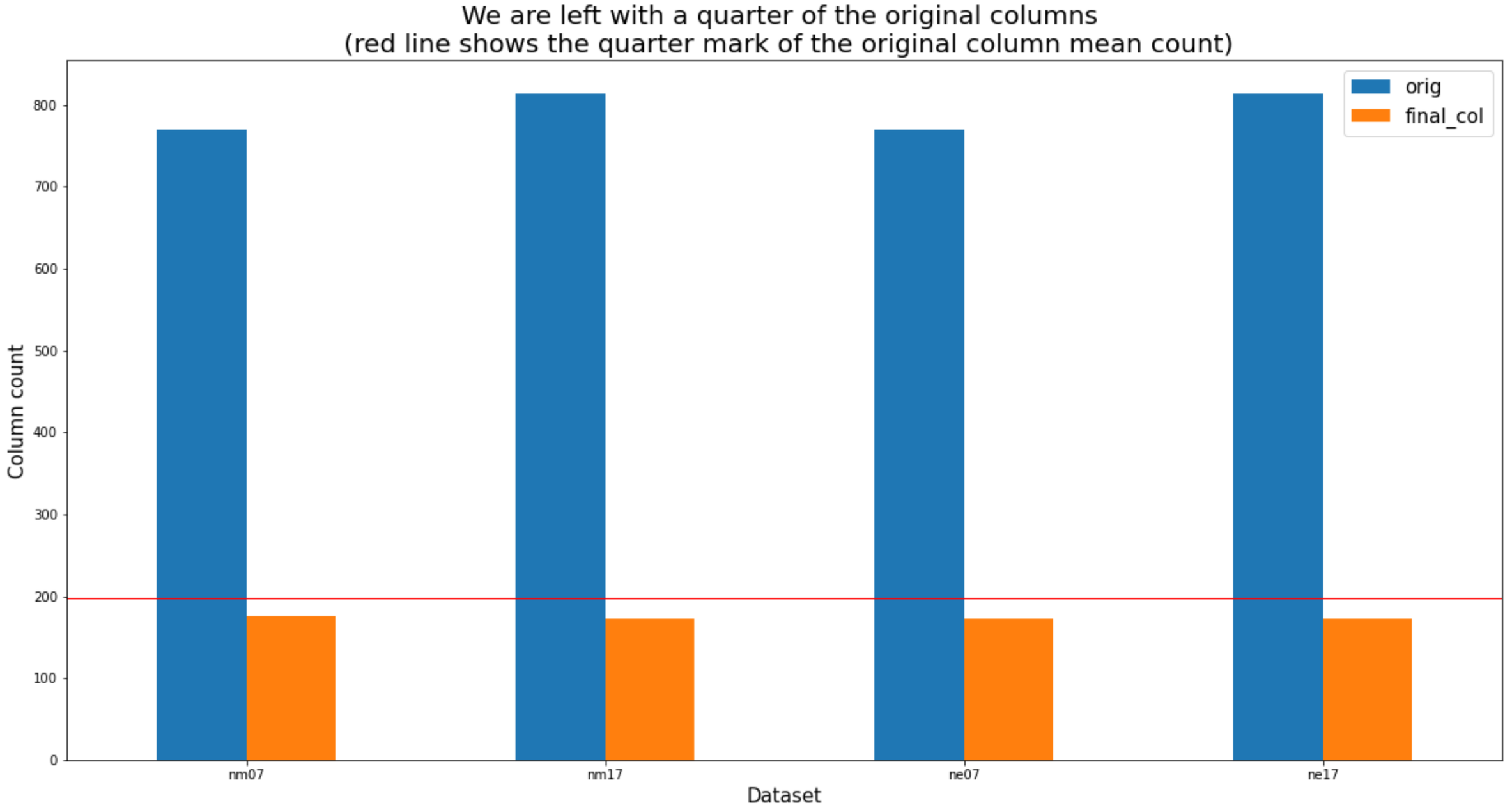




- Except for single moms as head of household, coupled with number of children.
- The majority are small households with 1 or 2 kids, not large families.
  - Note: 2007 had more and 2017 had less.

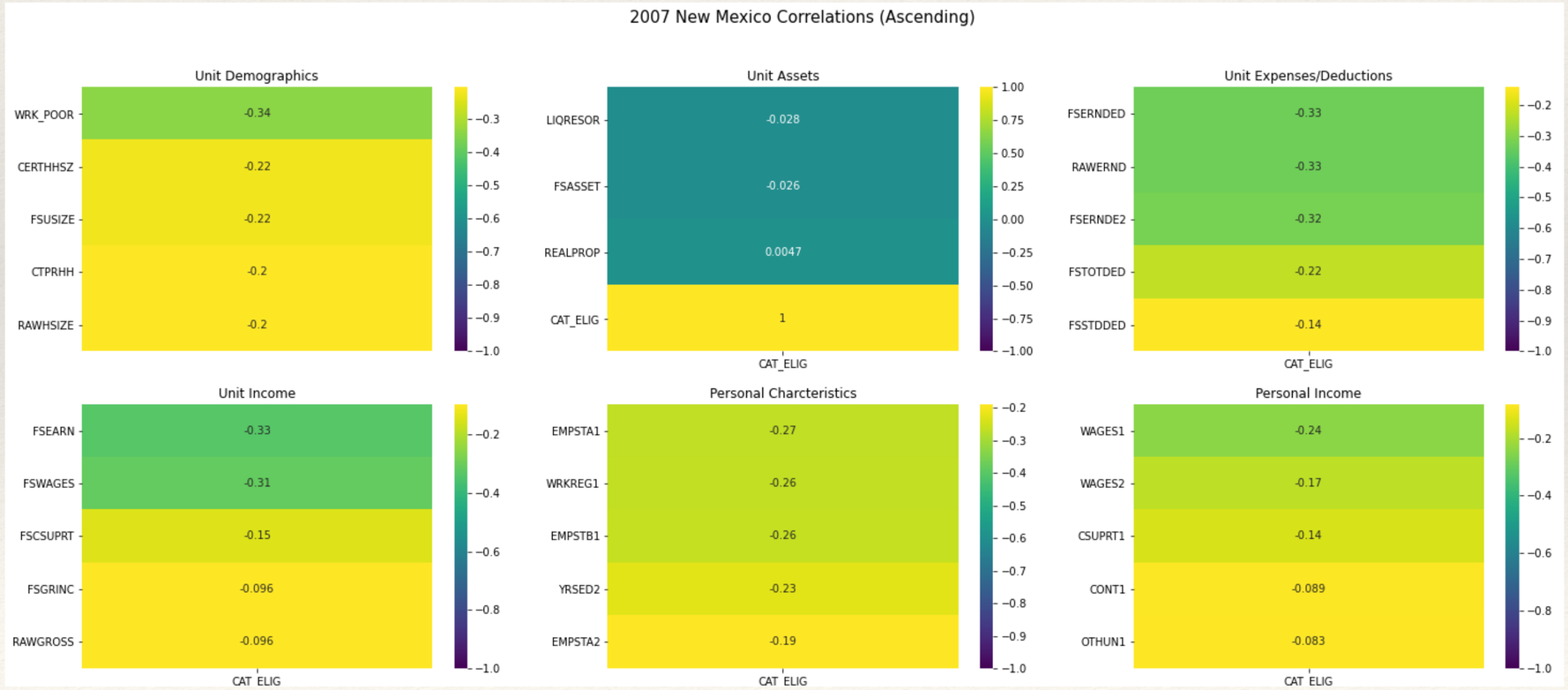
Narrowing down the data: High Nullity

- Most of the data had a high degree of nulls to them. So I broke the nulls into three points:
  1. Remove columns with ALL null values.
  2. Then drop columns with more than 50% nulls.
     - According to a paper called “The proportion of missing data should not be used to guide decisions on multiple imputation”, the authors test and conclude that the value of the data is more important than the amount of missing information.
  3. Lastly, I imputed nulls with the mean for the remainder using sklearn Simple Imputer.


We are left with a quarter of the original columns
(red line shows the quarter mark of the original column mean count)
orig
final_col
Column count
0
100
200
300
400
500
600
700
800
nm07
nm17
ne07
ne17
Dataset

Narrowing down the data: Correlations

- 2007 had a mix of eligible and not eligible applications for these two states, while 2017 did not. Therefore, a correlation to the target variable could only be run on 2007.
- The technical document included 6 sections of observations. My final dataset was the top 5 correlated features per section as a final set of columns. Ending in 32 features + the target column.

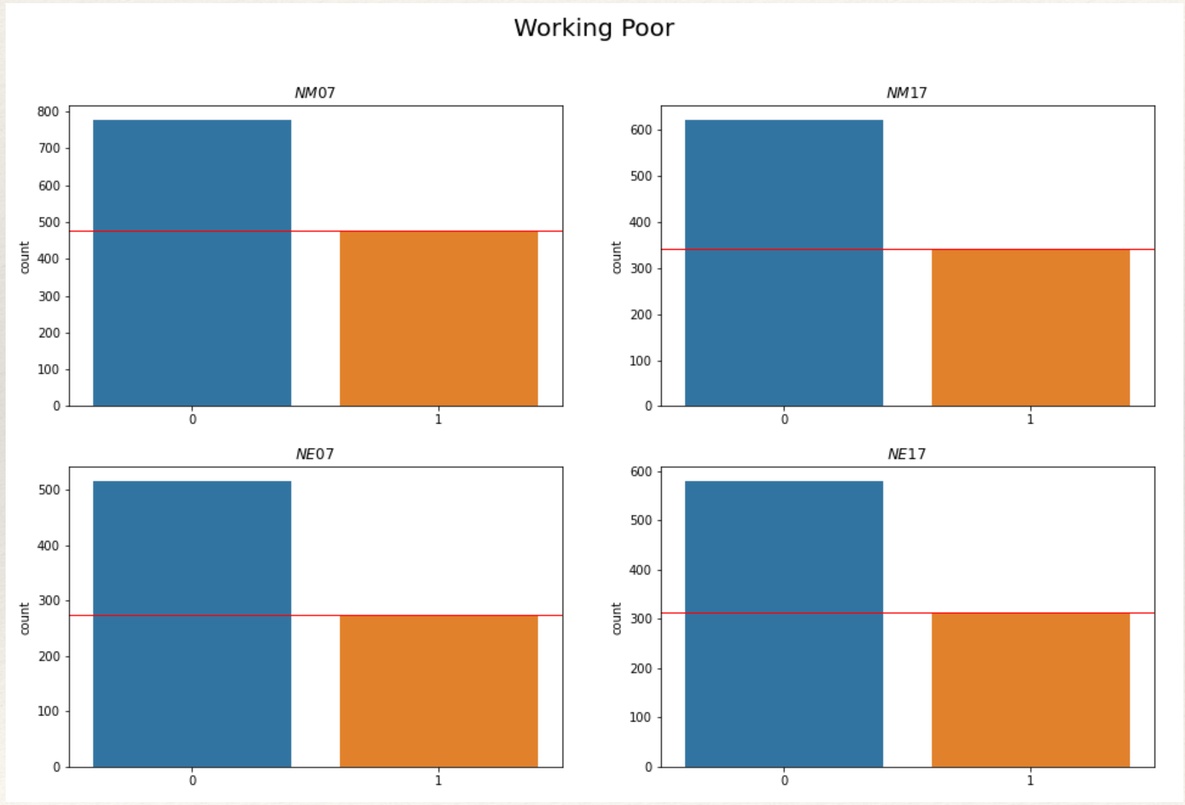


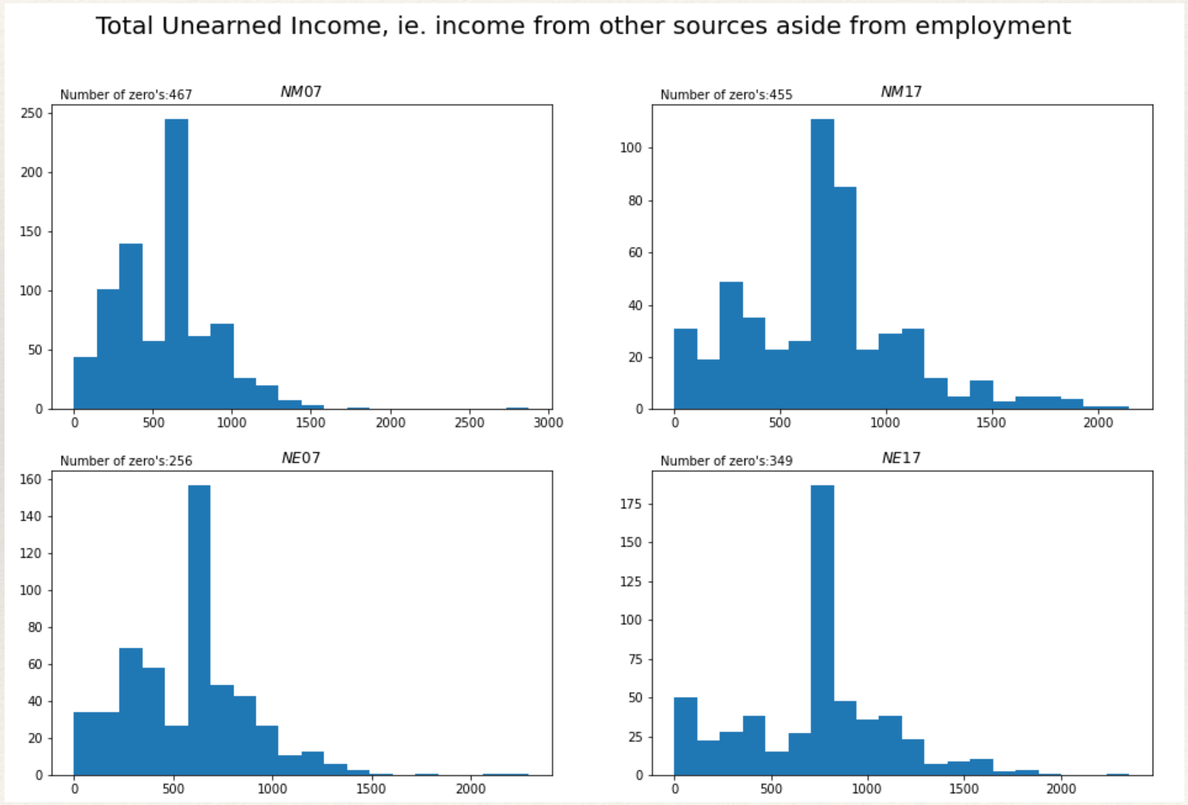

```
There are 32 features

{'CERTHHSZ',
 'FSASSET',
 'FSDIS',
 'FSEARN',
 'FSERNDE2',
 'FSERNDED',
 'FSGA',
 'FSGRINC',
 'FSNELDER',
 'FSNETINC',
 'FSNONCIT',
 'FSSLTDE2',
 'FSSLTDED',
 'FSSSI',
 'FSSTDDE2',
 'FSTANF',
 'FSTOTDE2',
 'FSTOTDED',
 'FSUNEARN',
 'FSUSIZE',
 'FSVEHAST',
 'FSWAGES',
 'HWGT',
 'LIQRESOR',
 'RAWERND',
 'RAWNET',
 'REALPROP',
 'SHELDED',
 'TANF_IND',
 'TPOV',
 'VEHICLEA',
 'WRK_POOR'}
```

Interesting Snapshots of the Data

1. New Mexico saw about 100 LESS working poor on SNAP in 2017. While Nebraska saw about 50 MORE in 2017.
2. In 2007, SNAP recipients were receiving less assistance from other welfare programs than in 2017.





The model

- I ran a number of different models. Random Forest and Gradient Boost performed the best.
- My feature selection process reduced enough noise to show only slight overfitting.
  - I added a PCA component to the test, and it greatly reduced the accuracy scores of all tests, thus supporting my feature selection was just right.
- I added a Bagging Classifier to further reduce the overfitting.

The model

- Initial model test comparison

name	cross_val_train	cross_val_test	test_recall	test_precision
LogReg	0.884334	0.862628	0.900262	0.903821
Decision Tree	0.907760	0.913823	0.943570	0.941099
Random Forest	0.942533	0.934300	0.965879	0.955844
Gradient Boost	0.940337	0.927474	0.950131	0.961487
Ada Boost	0.922035	0.920648	0.937008	0.939474
SVC	0.884700	0.854949	0.901575	0.928378
Naive Bayes	0.767204	0.741468	0.833333	0.908441

Final Model:

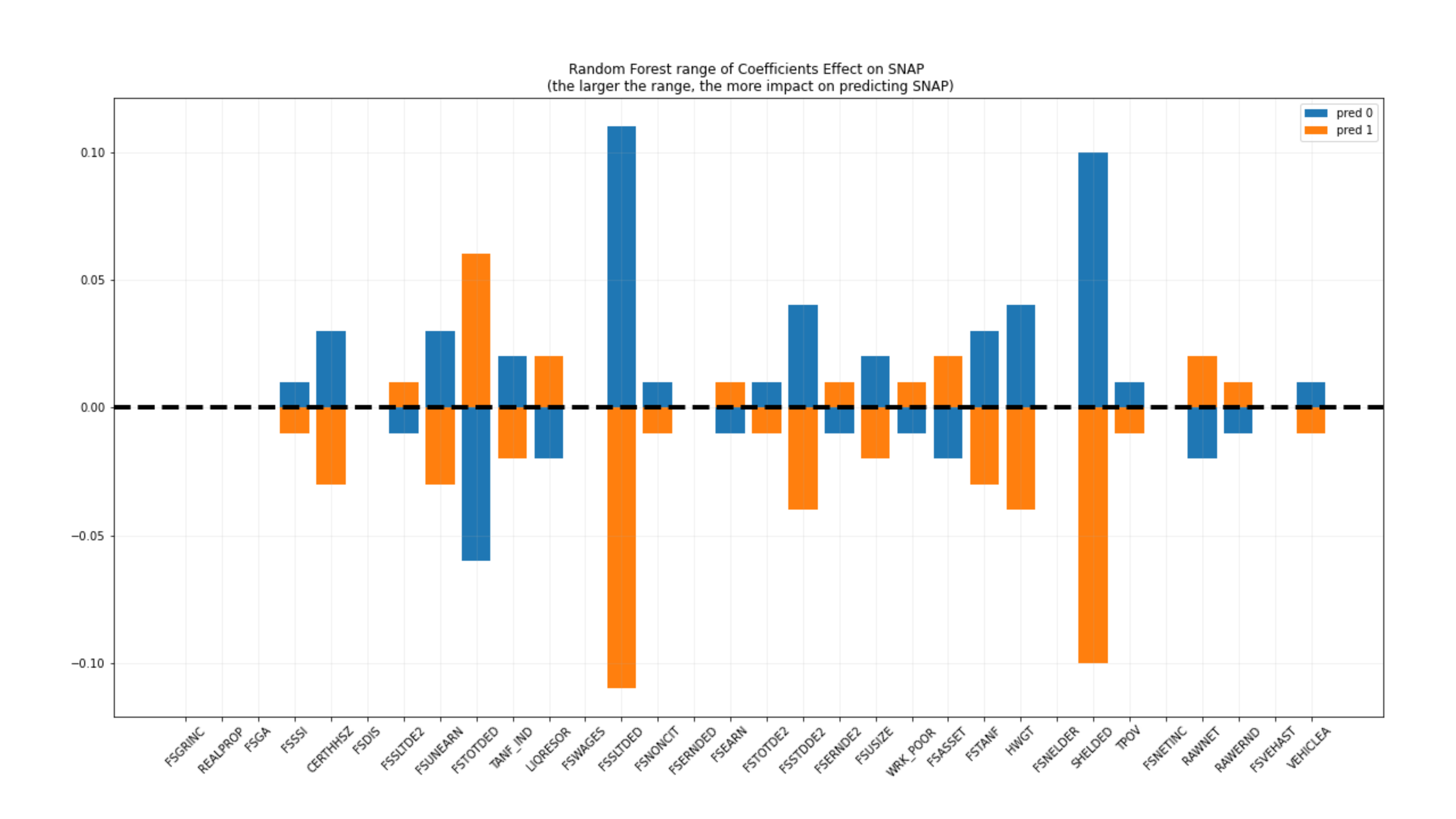
Voting Classifier with

Random Forest, Gradient Boost, and Bagging Classifier

CV best score: 95%

Interestingly, the best parameters indicated no bootstrapping on Random Forest, but yes to bootstrapping in the Bagging Classifier.

Also, the Bagging Classifier increased my recall and precision scores to 95%.

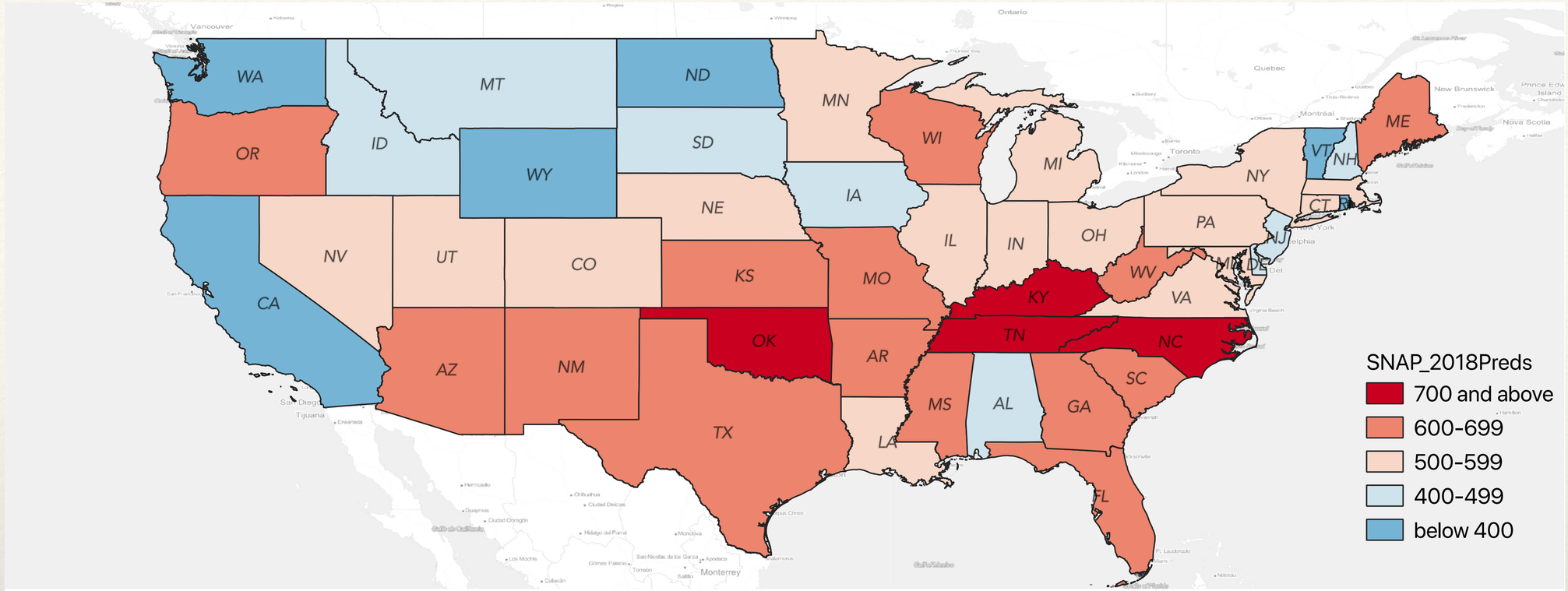

Random Forest range of Coefficients Effect on SNAP
(the larger the range, the more impact on predicting SNAP)
pred 0
pred 1
0.10
0.05
0.00
−0.05
−0.10
FSGRINC
REALPROP
FSGA
FSSSI
CERTHHSZ
FSDIS
FSSLTDE2
FSUNEARN
FSTOTDED
TANF_IND
LIQRESOR
FSWAGES
FSSLTDED
FSNONCIT
FSERNDED
FSEARN
FSTOTDE2
FSSTDDE2
FSERNDE2
FSUSIZE
WRK_POOR
FSASSET
FSTANF
HWGT
FSNELDER
SHELDED
TPOV
FSNETINC
RAWNET
RAWERND
FSVEHAST
VEHICLEA

Model Coefficients

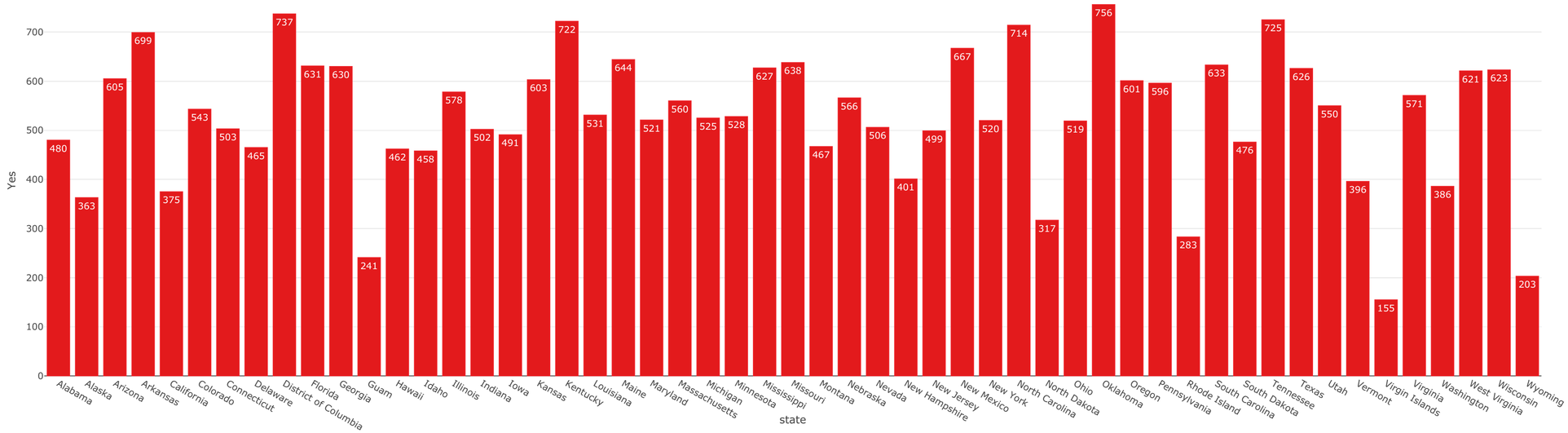
- I chose the models to test because I wanted an interpretable model.
- Random Forest uses *sklearn.treeinterpreter* to rate the impact of features on model prediction.
- **The top 4 variables deal with shelter and homelessness:**
  - FSSLTDED and SHELDED are indicators of how much someone is paying for their home.
  - FSTOTDED & FSSTDDE2 are deductions relating to housing.

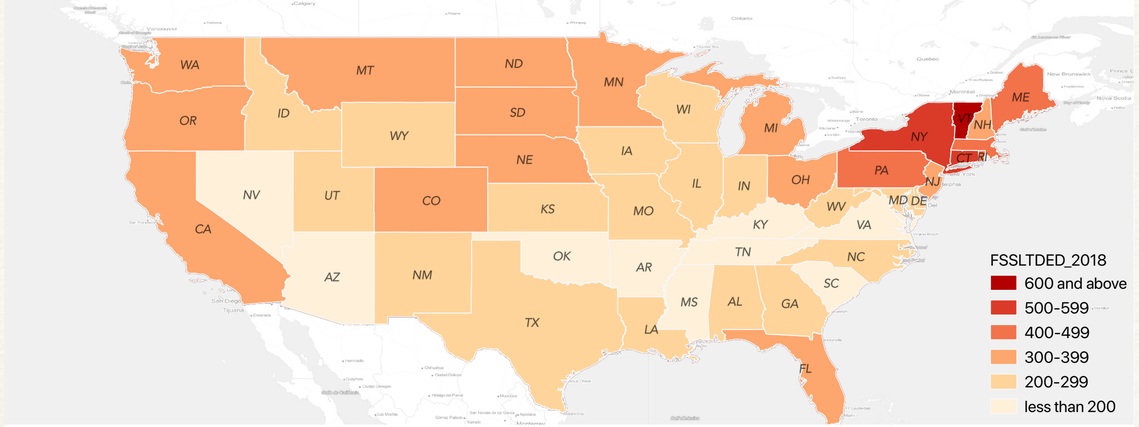
The top 15 impactful features:

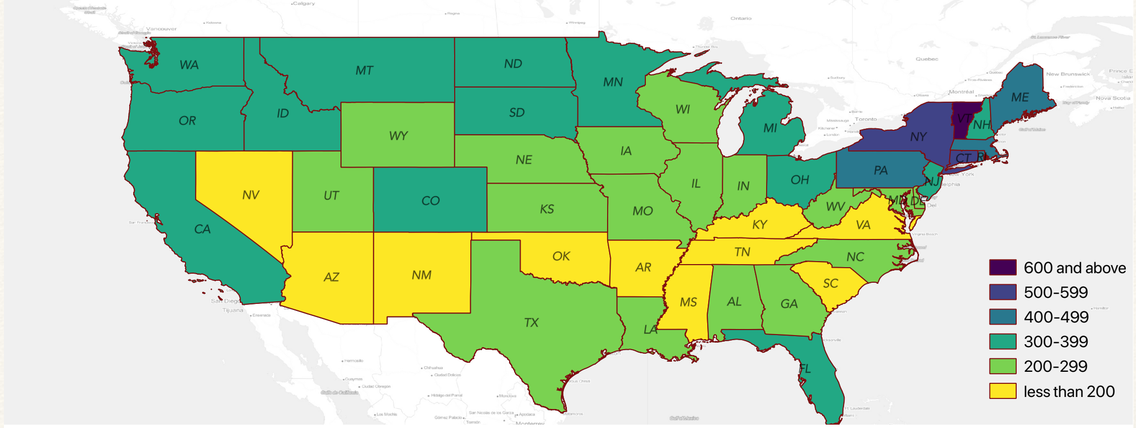
	feature	diff
0	FSSLTDED	0.22
1	SHELDED	0.20
2	FSTOTDED	0.12
3	FSSTDDE2	0.08
4	HWGT	0.08
5	CERTHHSZ	0.06
6	FSTANF	0.06
7	FSUNEARN	0.06
8	LIQRESOR	0.04
9	RAWNET	0.04
10	FSASSET	0.04
11	TANF_IND	0.04
12	FSUSIZE	0.04
13	VEHICLEA	0.02
14	FSSSI	0.02



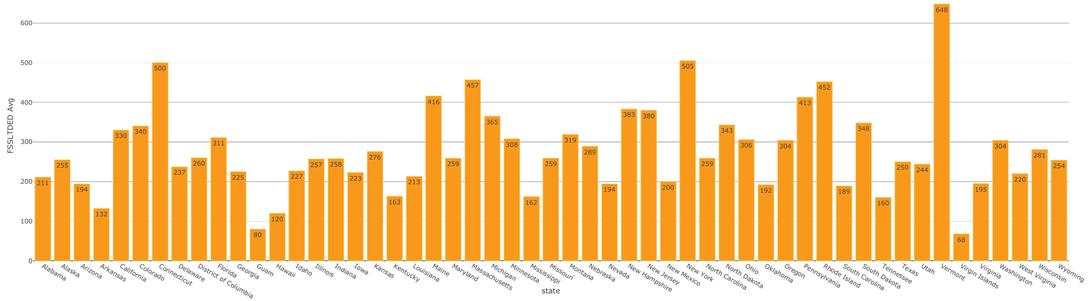
2018 Predicted SNAP Participants by State



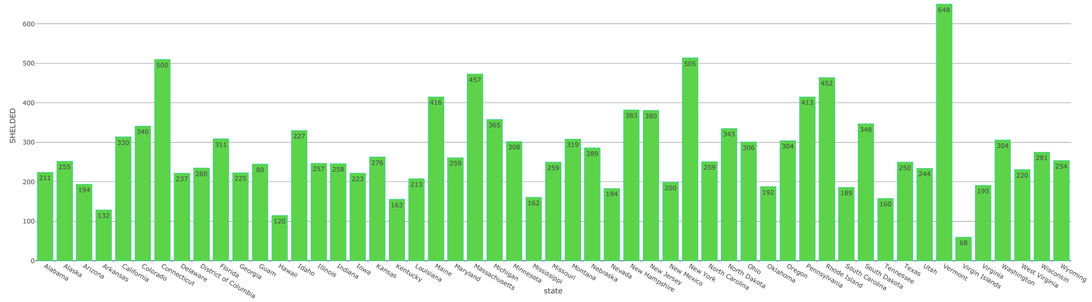


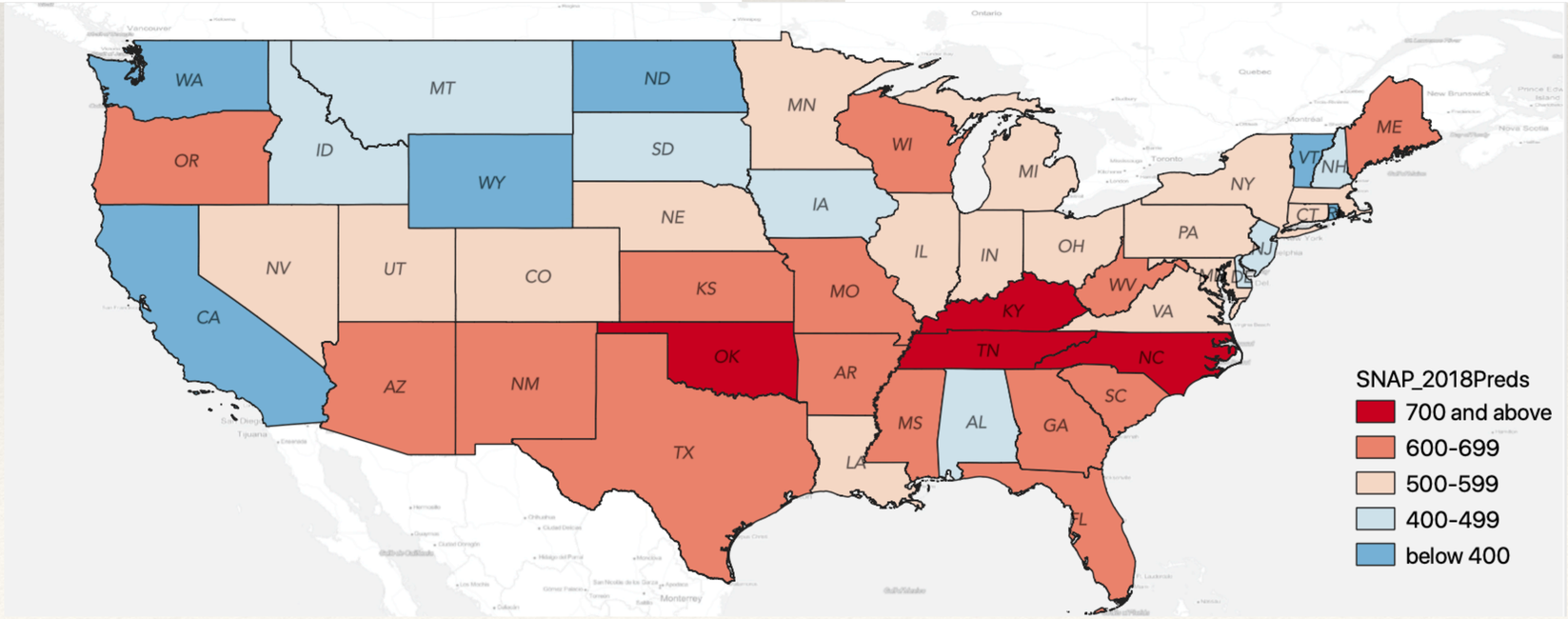


Calculated Excess Shelter Deduction: Average by State
(the taller the bar, the larger deduction allowed)



Reported Shelter Deduction: Average by State
(the taller the bar, the larger deduction allowed)





Next Steps…

- A geographic analysis that consists of access to housing resources such as HUD (which has a GIS page), food pantries and counties showing high levels of SNAP dependency. This would pinpoint areas where assistant could be targeted. Especially during COVID, targeting high need areas would be good way to direct tight resources.
- I would also add economic factors such as technologies in the area to see if it relates to swings in housing.
- Post this final analysis to an interactive dashboard for governments, charitable organizations and community activists.